Chapter 4

Introduction 208

Section 4.1 209
Sampling and Surveys

Section 4.2 234
Experiments

Section 4.3 266
Using Studies Wisely

Free Response AP® Problem, YAY! 275

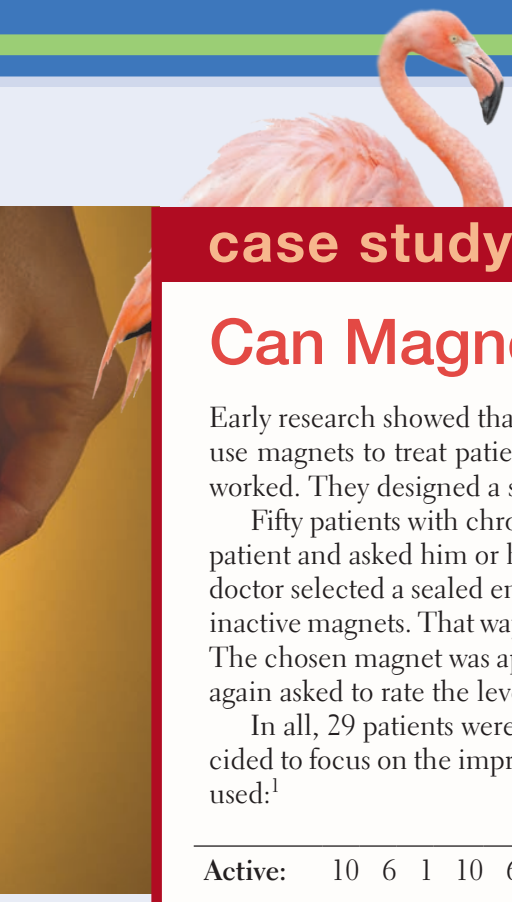
Chapter 4 Review 276

Chapter 4 Review Exercises 278

Chapter 4 AP® Statistics Practice Test 279

Cumulative AP® Practice Test 1 282

Designing Studies



case study

Can Magnets Help Reduce Pain?

Early research showed that magnetic fields affected living tissue in humans. Some doctors have begun to use magnets to treat patients with chronic pain. Scientists wondered whether this type of therapy really worked. They designed a study to find out.

Fifty patients with chronic pain were recruited for the study. A doctor identified a painful site on each patient and asked him or her to rate the pain on a scale from 0 (mild pain) to 10 (severe pain). Then, the doctor selected a sealed envelope containing a magnet at random from a box with a mixture of active and inactive magnets. That way, neither the doctor nor the patient knew which type of magnet was being used. The chosen magnet was applied to the site of the pain for 45 minutes. After "treatment," each patient was again asked to rate the level of pain from 0 to 10.

In all, 29 patients were given active magnets and 21 patients received inactive magnets. Scientists decided to focus on the improvement in patients' pain ratings. Here they are, grouped by the type of magnet used:[1]

Active:	10	6	1	10	6	8	5	5	6	8	7	8	7	6	4	4	7	10	6	10	6	5	5	1	0	0	0	0	1
Inactive:	4	3	5	2	1	4	1	0	0	1	0	0	0	0	0	0	0	1	0	0	1								

What do the data tell us about whether the active magnets helped reduce pain? By the end of the chapter, you'll be ready to interpret the results of this study.

Introduction

You can hardly go a day without hearing the results of a statistical study. Here are some examples:

- The National Highway Traffic Safety Administration (NHTSA) reports that seat belt use in passenger vehicles increased from 84% in 2011 to 86% in 2012.[2]

- According to a recent survey, U.S. teens aged 13 to 18 spend an average of 26.8 hours per week online. Although 59% of the teens said that posting personal information or photos online is unsafe, 62% said they had posted photos of themselves.[3]
- A recent study suggests that lack of sleep increases the risk of catching a cold.[4]
- For their final project, two AP® Statistics students showed that listening to music while studying decreased subjects' performance on a memory task.[5]

Can we trust these results? As you'll learn in this chapter, that depends on how the data were produced. Let's take a closer look at where the data came from in each of these studies.

Each year, the NHTSA conducts an *observational study* of seat belt use in vehicles. The NHTSA sends trained observers to record the actual behavior of people in vehicles at randomly selected locations across the country. The idea of an observational study is simple: you can learn a lot just by watching. Or by asking a few questions, as in the survey of teens' online habits. Harris Interactive conducted this survey using a "representative sample" of 655 U.S. 13- to 18-year-olds. Both of these studies use information from a *sample* to draw conclusions about some larger *population*. Section 4.1 examines the issues involved in sampling and surveys.

In the sleep and catching a cold study, 153 volunteers took part. They answered questions about their sleep habits over a two-week period. Then, researchers gave them a virus and waited to see who developed a cold. This was a complicated observational study. Compare this with the *experiment* performed by the AP® Statistics students. They recruited 30 students and divided them into two groups of 15 by drawing names from a hat. Students in one group tried to memorize a list of words while listening to music. Students in the other group tried to memorize the same list of words while sitting in silence. Section 4.2 focuses on designing experiments.

The goal of many statistical studies is to show that changes in one variable *cause* changes in another variable. In Section 4.3, we'll look at why establishing causation is so difficult, especially in observational studies. We'll also consider some of the ethical issues involved in planning and conducting a study.

Here's an Activity that gives you a preview of what lies ahead.

ACTIVITY | See no evil, hear no evil?

MATERIALS:

Two index cards, each with 10 distinct numbers from 00 to 99 written on it (prepared by your teacher); clock, watch, or stopwatch to measure 30 seconds; and a coin for each pair of students

Confucius said, "I hear and I forget. I see and I remember. I do and I understand." Do people really remember what they see better than what they hear?[6] In this Activity, you will perform an experiment to try to find out.

1. Divide the class into pairs of students by drawing names from a hat.
2. Your teacher will give each pair two index cards with 10 distinct numbers from 00 to 99 on them. *Do not look at the numbers* until it is time for you to do the experiment.
3. Flip a coin to decide which of you is Student 1 and which is Student 2. Shuffle the index cards and deal one face down to each partner.
4. Student 1 will be the first to attempt a memory task while Student 2 keeps time.

Directions: Study the numbers on the index card for 30 seconds. Then turn the card over. Recite the alphabet aloud (A, B, C, and so on). Then tell your partner what you think the numbers on the card are. *You may not say more than 10 numbers!* Student 2 will record how many numbers you recalled correctly.

5. Now it's Student 2's turn to do a memory task while Student 1 records the data.

Directions: Your partner will read the numbers on your index card aloud three times slowly. Next, you will recite the alphabet aloud (A, B, C, and so on) and then tell your partner what you think the numbers on the card are. *You may not say more than 10 numbers!* Student 1 will record how many numbers you recalled correctly.

6. Your teacher will scale and label axes on the board for parallel dotplots of the results. Plot how many numbers you remembered correctly on the appropriate graph.
7. Did students in your class remember numbers better when they saw them or when they heard them? Give appropriate evidence to support your answer.
8. Based on the results of this experiment, can we conclude that people in general remember better when they see than when they hear? Why or why not?

4.1 Sampling and Surveys

WHAT YOU WILL LEARN By the end of the section, you should be able to:

- Identify the population and sample in a statistical study.
- Identify voluntary response samples and convenience samples. Explain how these sampling methods can lead to bias.
- Describe how to obtain a random sample using slips of paper, technology, or a table of random digits.
- Distinguish a simple random sample from a stratified random sample or cluster sample. Give the advantages and disadvantages of each sampling method.
- Explain how undercoverage, nonresponse, question wording, and other aspects of a sample survey can lead to bias.

Suppose we want to find out what percent of young drivers in the United States text while driving. To answer the question, we will survey 16- to 20-year-olds who live in the United States and drive. Ideally, we would ask them all (take a **census**). But contacting every driver in this age group wouldn't be practical: it would take too much time and cost too much money. Instead, we put the question to a **sample** chosen to represent the entire **population** of young drivers.

DEFINITION: Population, census, and sample

The **population** in a statistical study is the entire group of individuals we want information about. A **census** collects data from every individual in the population.

A **sample** is a subset of individuals in the population from which we actually collect data.

The distinction between population and sample is basic to statistics. To make sense of any sample result, you must know what population the sample represents. Here's an example that illustrates this distinction and also introduces some major uses of sampling.

EXAMPLE Sampling Hardwood and Humans

Populations and samples

PROBLEM: Identify the population and the sample in each of the following settings.

(a) A furniture maker buys hardwood in large batches. The supplier is supposed to dry the wood before shipping (wood that isn't dry won't hold its size and shape). The furniture maker chooses five pieces of wood from each batch and tests their moisture content. If any piece exceeds 12% moisture content, the entire batch is sent back.

(b) Each week, the Gallup Poll questions a sample of about 1500 adult U.S. residents to determine national opinion on a wide variety of issues.

SOLUTION:

(a) The population is all the pieces of hardwood in a batch. The sample is the five pieces of wood that are selected from that batch and tested for moisture content.

(b) Gallup's population is all adult U.S. residents. Their sample is the 1500 adults who actually respond to the survey questions.

For Practice *Try Exercise* 1

The Idea of a Sample Survey

We often draw conclusions about a whole population on the basis of a sample. Have you ever tasted a sample of ice cream and ordered a cone if the sample tastes good? Because ice cream is fairly uniform, the single taste represents the whole. Choosing a representative sample from a large and varied population (like all young U.S. drivers) is not so easy. The first step in planning a **sample survey** is to say exactly *what population* we want to describe. The second step is to say exactly *what we want to measure*, that is, to give exact definitions of our variables.

We reserve the term "sample survey" for studies that use an organized plan to choose a sample that represents some specific population, like the pieces of hardwood and the U.S. adults in the previous example. By our definition, the population in a sample survey can consist of people, animals, or things. Some people use the terms "survey" or "sample survey" to refer only to studies in which people are asked one or more questions, like the Gallup Poll of the last example. We'll avoid this more restrictive terminology.

EXAMPLE

How Does the Current Population Survey Work?

A sample survey

One of the most important government sample surveys in the United States is the monthly Current Population Survey (CPS). The CPS contacts about 60,000 households each month. It produces the monthly unemployment rate and lots of other economic and social information. To measure unemployment, we must first specify the population we want to describe. The CPS defines its population as all U.S. residents (legal or not) 16 years of age and over who are civilians and are not in an institution such as a prison. The unemployment rate announced in the news refers to this specific population.

What does it mean to be "unemployed"? Someone who is not looking for work—for example, a full-time student—should not be called unemployed just because she is not working for pay. If you are chosen for the CPS sample, the interviewer first asks whether you are available to work and whether you actually looked for work in the past four weeks. If not, you are neither employed nor unemployed—you are not in the labor force.

If you are in the labor force, the interviewer goes on to ask about employment. If you did any work for pay or in your own business during the week of the survey, you are employed. If you worked at least 15 hours in a family business without pay, you are employed. You are also employed if you have a job but didn't work because of vacation, being on strike, or other good reason. An unemployment rate of 9.7% means that 9.7% of the sample was unemployed, using the exact CPS definitions of both "labor force" and "unemployed."

The final step in planning a sample survey is to decide how to choose a sample from the population. Let's take a closer look at some good and not-so-good sampling methods.

How to Sample Badly

The sampling method that yields a convenience sample is called *convenience sampling*. Other sampling methods are named in similarly obvious ways!

Suppose we want to know how long students at a large high school spent doing homework last week. We might go to the school library and ask the first 30 students we see about their homework time. The sample we get is known as a **convenience sample.**

DEFINITION: Convenience sample

Choosing individuals from the population who are easy to reach results in a **convenience sample**.

Convenience sampling often produces unrepresentative data. Consider our sample of 30 students from the school library. It's unlikely that this convenience sample accurately represents the homework habits of all students at the high school. In fact, if we were to repeat this sampling process again and again, we would almost always overestimate the average homework time in the population. Why? Because students who hang out in the library tend to be more studious. This is **bias:** using a method that favors some outcomes over others.

DEFINITION: Bias

The design of a statistical study shows **bias** if it would consistently underestimate or consistently overestimate the value you want to know.

AP® EXAM TIP If you're asked to describe how the design of a study leads to bias, you're expected to do two things: (1) identify a problem with the design, and (2) explain how this problem would lead to an underestimate or overestimate. Suppose you were asked, "Explain how using your statistics class as a sample to estimate the proportion of all high school students who own a graphing calculator could result in bias." You might respond, "This is a convenience sample. It would probably include a much higher proportion of students with a graphing calculator than in the population at large because a graphing calculator is required for the statistics class. So this method would probably lead to an overestimate of the actual population proportion."

Bias is not just bad luck in one sample. It's the result of a bad study design that will consistently miss the truth about the population in the same way. Convenience samples are almost guaranteed to show bias. So are **voluntary response samples**.

Voluntary response samples are also known as *self-selected samples*.

DEFINITION: Voluntary response sample

A **voluntary response sample** consists of people who choose themselves by responding to a general invitation.

The Internet brings voluntary response samples to the computer nearest you. Visit www.misterpoll.com to become part of the sample in any of dozens of online polls. As the site says, "None of these polls are 'scientific,' but do represent the collective opinion of everyone who participates." Unfortunately, such polls don't tell you anything about the views of the population.

Call-in, text-in, write-in, and many Internet polls rely on voluntary response samples. *People who choose to participate in such surveys are usually not representative of some larger population of interest*. Voluntary response samples attract people who feel strongly about an issue, and who often share the same opinion. That leads to bias.

EXAMPLE

Illegal Immigration

Online polls

Former CNN commentator Lou Dobbs doesn't like illegal immigration. One of his shows was largely devoted to attacking a proposal to offer driver's licenses to illegal immigrants. During the show, Mr. Dobbs invited his viewers to go to loudobbs.com to vote on the question "Would you be more or less likely to vote for a presidential candidate who supports giving driver's licenses to illegal aliens? The result: 97% of the 7350 people who voted by the end of the show said, "Less likely."

PROBLEM: What type of sample did Mr. Dobbs use in his poll? Explain how this sampling method could lead to bias in the poll results.

SOLUTION: Mr. Dobbs used a voluntary response sample: people chose to go online and respond. Those who voted were viewers of Mr. Dobbs's program, which means that they are likely to support his views. The 97% poll result is probably an extreme overestimate of the percent of people in the population who would be less likely to support a presidential candidate with this position.

For Practice *Try Exercise* 9

CHECK YOUR UNDERSTANDING

For each of the following situations, identify the sampling method used. Then explain how the sampling method could lead to bias.

1. A farmer brings a juice company several crates of oranges each week. A company inspector looks at 10 oranges from the top of each crate before deciding whether to buy all the oranges.

2. The ABC program *Nightline* once asked whether the United Nations should continue to have its headquarters in the United States. Viewers were invited to call one telephone number to respond "Yes" and another for "No." There was a charge for calling either number. More than 186,000 callers responded, and 67% said "No."

How to Sample Well: Simple Random Sampling

In convenience sampling, the researcher chooses easy-to-reach members of the population. In voluntary response sampling, people decide whether to join the sample. Both sampling methods suffer from bias due to personal choice. The best way to avoid this problem is to let chance choose the sample. That's the idea of **random sampling**.

DEFINITION: Random Sampling

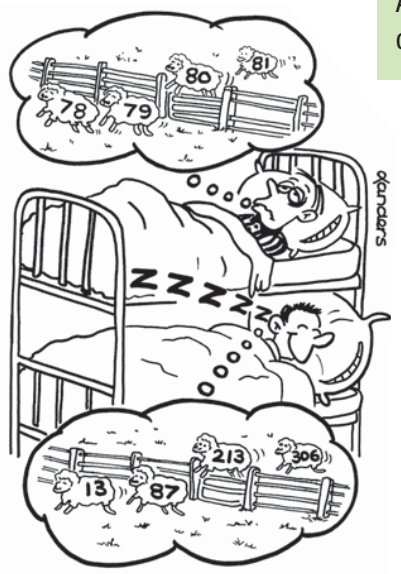
Random sampling involves using a chance process to determine which members of a population are included in the sample.

In everyday life, some people use the word "random" to mean haphazard, as in "that's so random." In statistics, random means "due to chance." Don't say that a sample was chosen at random if a chance process wasn't used to select the individuals.

The easiest way to choose a random sample of n people is to write their names on identical slips of paper, put the slips in a hat, mix them well, and pull out slips one at a time until you have n of them. An alternative would be to give each member of the population a distinct number and to use the "hat method" with these numbers instead of people's names. Note that this version would work just as well if the population consisted of animals or things. The resulting sample is called a **simple random sample,** or SRS for short.

DEFINITION: Simple Random Sample (SRS)

A **simple random sample** (SRS) of size n is chosen in such a way that every group of n individuals in the population has an equal chance to be selected as the sample.



Statisticians fall asleep faster by taking a random sample of sheep.

An SRS gives every possible sample of the desired size an equal chance to be chosen. It also gives each member of the population an equal chance to be included in the sample. Picture drawing 20 slips (the sample) from a hat containing 200 identical slips (the population). Any 20 slips have the same chance as any other 20 to be chosen. Also, each slip has a 1-in-10 chance (20/200) of being selected.

Some other random sampling methods give each member of the population, but not each sample, an equal chance. We'll look at some of these later.

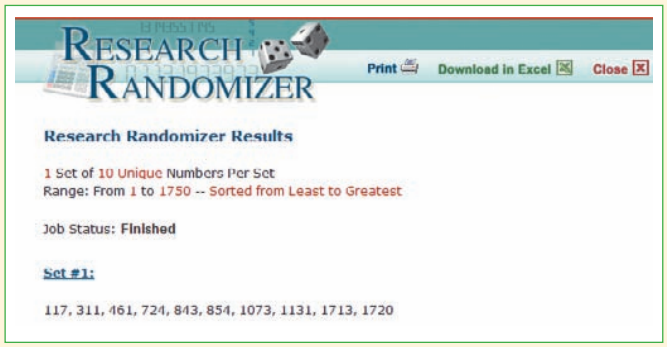
How to Choose a Simple Random Sample The hat method won't work well if the population is large. Imagine trying to take a simple random sample of 1000 U.S. adults! In practice, most people use random numbers generated by technology to choose samples.

EXAMPLE Teens on the Internet

Choosing an SRS with technology

The principal at Canyon del Oro High School in Arizona wants student input about limiting access to certain Internet sites on the school's computers. He asks the AP® Statistics teacher, Mr. Tabor, to select a "representative sample" of 10 students. Mr. Tabor decides to take an SRS from the 1750 students enrolled this year.

He gets an alphabetical roster from the registrar's office, and numbers the students from 1 to 1750. Then Mr. Tabor uses the random number generator at www.randomizer.org to choose 10 distinct numbers between 1 and 1750:



The 10 students on the roster that correspond to the chosen numbers will be on the principal's committee.

This example highlights the steps in **choosing a simple random sample with technology**.

It is standard practice to use *n* for the sample size and *N* for the population size.

CHOOSING AN SRS WITH TECHNOLOGY

Step 1: Label. Give each individual in the population a distinct numerical label from 1 to *N*.

Step 2: Randomize. Use a random number generator to obtain *n* different integers from 1 to *N*.

You can also use a graphing calculator to choose an SRS.

10. TECHNOLOGY CORNER CHOOSING AN SRS

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

Let's use a graphing calculator to select an SRS of 10 students from the Canyon del Oro High School roster.

1. Check that your calculator's random number generator is working properly.

TI-83/84

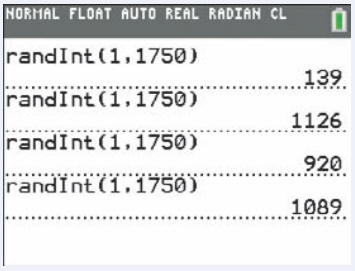
- Press [MATH], then select `PRB` and `randInt(`. Complete the command `randInt(1,1750)` and press [ENTER].

TI-89

- Press [CATALOG], then [F3] (Flash Apps) and choose `randInt(`. Complete the command `TIStat.randInt(1,1750)` and press [ENTER].

Compare your results with those of your classmates. If several students got the same number, you'll need to seed your calculator's random integer generator with different numbers before you proceed. Directions for doing this are given in the *Annotated Teacher's Edition*.

2. Randomly generate 10 distinct numbers from 1 to 1750.
   Do `randInt(1,1750)` again. Keep pressing [ENTER] until you have chosen 10 different labels.



Note: If you have a TI-83/84 with OS 2.55 or later, you can use the command `RandIntNoRep(1,1750)` to sort the numbers from 1 to 1750 in random order. The first 10 numbers listed give the labels of the chosen students.

If you don't have technology handy, you can use a table of random digits to choose an SRS. We have provided a table of random digits at the back of the book (Table D). Here is an excerpt.

Table D Random digits

LINE								
101	19223	95034	05756	28713	96409	12531	42544	82853
102	73676	47150	99400	01927	27754	42648	82425	36290
103	45467	71709	77558	00095	32863	29485	82226	90056

You can think of this table as the result of someone putting the digits 0 to 9 in a hat, mixing, drawing one, replacing it, mixing again, drawing another, and so on. The digits have been arranged in groups of five within numbered rows to make the table easier to read. The groups and rows have no special meaning—Table D is just a long list of randomly chosen digits. As with technology, there are two steps in using Table D to choose a random sample.

HOW TO CHOOSE AN SRS USING TABLE D

Step 1: Label. Give each member of the population a numerical label with the same number of digits. Use as few digits as possible.

Step 2: Randomize. Read consecutive groups of digits of the appropriate length from left to right across a line in Table D. Ignore any group of digits that wasn't used as a label or that duplicates a label already in the sample. Stop when you have chosen n different labels.

Your sample contains the individuals whose labels you find.

Always use the shortest labels that will cover your population. For instance, you can label up to 100 individuals with two digits: 01, 02, . . . , 99, 00. As standard practice, we recommend that you begin with label 1 (or 01 or 001 or 0001, as needed). Reading groups of digits from the table gives all individuals the same chance to be chosen because all labels of the same length have the same chance

to be found in the table. For example, any pair of digits in the table is equally likely to be any of the 100 possible labels 01, 02, . . . , 99, 00. Here's an example that shows how this process works.

EXAMPLE Spring Break!

Choosing an SRS with Table D

The school newspaper is planning an article on family-friendly places to stay over spring break at a nearby beach town. The editors intend to call 4 randomly chosen hotels to ask about their amenities for families with children. They have an alphabetized list of all 28 hotels in the town.

PROBLEM: Use Table D at line 130 to choose an SRS of 4 hotels for the editors to call.

SOLUTION: We'll use the two-step process for selecting an SRS using Table D.

Step 1: Label. Two digits are needed to label the 28 hotels. We have added labels 01 to 28 to the alphabetized list of hotels below.

01	Aloha Kai	08	Captiva	15	Palm Tree	22	Sea Shell
02	Anchor Down	09	Casa del Mar	16	Radisson	23	Silver Beach
03	Banana Bay	10	Coconuts	17	Ramada	24	Sunset Beach
04	Banyan Tree	11	Diplomat	18	Sandpiper	25	Tradewinds
05	Beach Castle	12	Holiday Inn	19	Sea Castle	26	Tropical Breeze
06	Best Western	13	Lime Tree	20	Sea Club	27	Tropical Shores
07	Cabana	14	Outrigger	21	Sea Grape	28	Veranda

Step 2: Randomize. To use Table D, start at the left-hand side of line 130 and read two-digit groups. Skip any groups that aren't between 01 and 28, as well as any repeated groups. Continue until you have chosen four hotels. Here is the beginning of line 130:

69051 64817 87174 09517 84534 06489 87201 97245

The first 10 two-digit groups are

69	05	16	48	17	87	17	40	95	17
Skip	✓	✓	Skip	✓	Skip	Skip	Skip	Skip	Skip
Too big			Too big		Too big	Repeat	Too big	Too big	Repeat

We skip 5 of these 10 groups because they are too high (over 28) and 2 because they are repeats (both 17s). The hotels labeled 05, 16, and 17 go into the sample. We need one more hotel to complete the sample. Continuing along line 130:

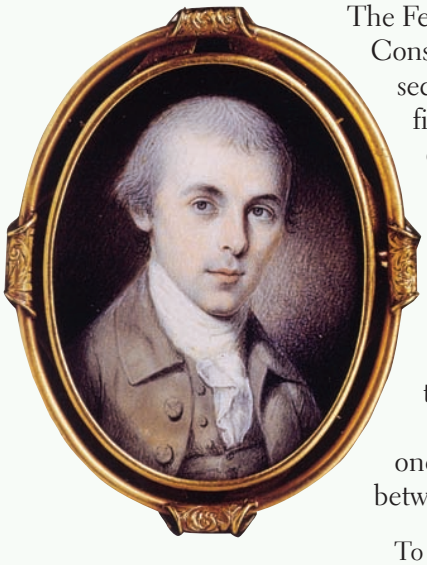
84	53	40	64	89	87	20
Skip	Skip	Skip	Skip	Skip	Skip	✓
Too big	Too big	Too big	Too big	Too big	Too big	

Our SRS of 4 hotels for the editors to contact is 05 Beach Castle, 16 Radisson, 17 Ramada, and 20 Sea Club.

For Practice *Try Exercise* 11

We can trust results from an SRS, as well as from other types of random samples that we will meet later, because the use of impersonal chance avoids bias. The following activity shows why random sampling is so important.

ACTIVITY | Who Wrote the Federalist Papers?



The Federalist Papers are a series of 85 essays supporting the ratification of the U.S. Constitution. At the time they were published, the identity of the authors was a secret known to just a few people. Over time, however, the authors were identified as Alexander Hamilton, James Madison, and John Jay. The authorship of 73 of the essays is fairly certain, leaving 12 in dispute. However, thanks in some part to statistical analysis,[7] most scholars now believe that the 12 disputed essays were written by Madison alone or in collaboration with Hamilton.[8]

There are several ways to use statistics to help determine the authorship of a disputed text. One method is to estimate the average word length in a disputed text and compare it to the average word lengths of works where the authorship is not in dispute.

The following passage is the opening paragraph of Federalist Paper #51,[9] one of the disputed essays. The theme of this essay is the separation of powers between the three branches of government.

> To what expedient, then, shall we finally resort, for maintaining in practice the necessary partition of power among the several departments, as laid down in the Constitution? The only answer that can be given is, that as all these exterior provisions are found to be inadequate, the defect must be supplied, by so contriving the interior structure of the government as that its several constituent parts may, by their mutual relations, be the means of keeping each other in their proper places. Without presuming to undertake a full development of this important idea, I will hazard a few general observations, which may perhaps place it in a clearer light, and enable us to form a more correct judgment of the principles and structure of the government planned by the convention.

1. Choose 5 words from this passage. Count the number of letters in each of the words you selected, and find the average word length.

2. Your teacher will draw and label a horizontal axis for a class dotplot. Plot the average word length you obtained in Step 1 on the graph.

3. Use a table of random digits or a random number generator to select a simple random sample of 5 words from the 130 words in the opening passage. Count the number of letters in each of the words you selected, and find the average word length.

4. Your teacher will draw and label another horizontal axis with the same scale for a comparative class dotplot. Plot the average word length you obtained in Step 3 on the graph.

5. How do the dotplots compare? Can you think of any reasons why they might be different? Discuss with your classmates.

Other Random Sampling Methods

The basic idea of sampling is straightforward: take an SRS from the population and use your sample results to gain information about the population. Unfortunately, it's usually difficult to get an SRS from the population of interest. Imagine trying to get a simple random sample of all the batteries produced in one day at a factory. Or an SRS of all U.S. high school students. In either case, it would be difficult to obtain an accurate list of the population from which to draw the sample. It would also be very time-consuming to collect data from each individual that's randomly selected. Sometimes, there are also statistical advantages to using more complex sampling methods.

One of the most common alternatives to an SRS involves sampling groups (**strata**) of similar individuals within the population separately. Then these separate "subsamples" are combined to form one **stratified random sample**.

Stratum is singular. Strata are plural.

DEFINITION: Stratified random sample and strata

To get a **stratified random sample**, start by classifying the population into groups of similar individuals, called **strata**. Then choose a separate SRS in each stratum and combine these SRSs to form the sample.

Choose the strata based on facts known before the sample is taken. For example, in a study of sleep habits on school nights, the population of students in a large high school might be divided into freshman, sophomore, junior, and senior strata. In a pre-election poll, a population of election districts might be divided into urban, suburban, and rural strata. Stratified random sampling works best when the individuals within each stratum are similar with respect to what is being measured and when there are large differences between strata. The following Activity makes this point clear.

ACTIVITY | Sampling sunflowers

MATERIALS:
Calculator for each student

A British farmer grows sunflowers for making sunflower oil. Her field is arranged in a grid pattern, with 10 rows and 10 columns as shown in the figure on the next page. Irrigation ditches run along the top and bottom of the field. The farmer would like to estimate the number of healthy plants in the field so she can project how much money she'll make from selling them. It would take too much time to count the plants in all 100 squares, so she'll accept an estimate based on a sample of 10 squares.

1. Use Table D or technology to take a simple random sample of 10 grid squares. Record the location (for example, B6) of each square you select.

2. This time, you'll take a stratified random sample using the *rows* as strata. Use Table D or technology to randomly select one square from each (horizontal) row. Record the location of each square—for example, Row 1: G, Row 2: B, and so on.

	A	B	C	D	E	F	G	H	I	J
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										

3. Now, take a stratified random sample using the *columns* as strata. Use Table D or technology to randomly select one square from each (vertical) column. Record the location of each square—for example, Column A: 4, Column B: 1, and so on.

4. The table on page N/DS-5 in the back of the book gives the actual number of sunflowers in each grid square. Use the information provided to calculate your estimate of the mean number of sunflowers per square for each of your samples in Steps 1, 2, and 3.

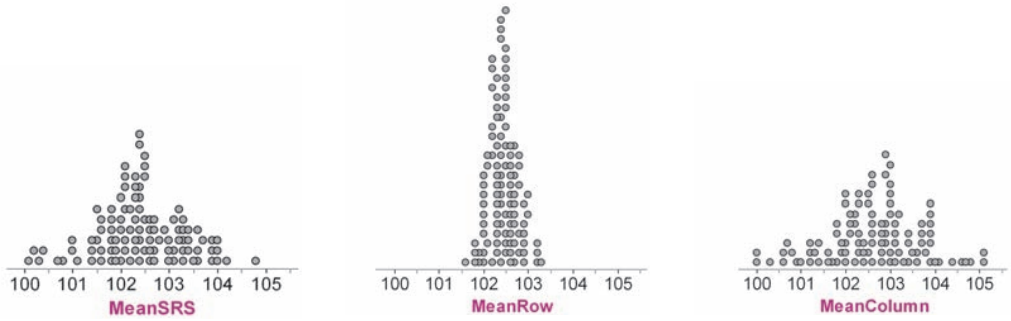
5. Make comparative dotplots showing the mean number of sunflowers obtained using the three different sampling methods for all members of the class. Describe any similarities and differences you see.

6. Your teacher will provide you with the mean number of sunflowers in the population of all 100 grid squares in the field. How did the three sampling methods do?

The dotplots below show the mean number of healthy plants in 100 samples using each of the three sampling methods in the Activity: simple random sampling, stratified random sampling with rows of the field as strata, and stratified random sampling with columns of the field as strata. Notice that all three distributions are centered at about 102.5, the true mean number of healthy plants in all squares of the field. That makes sense because random sampling yields accurate estimates of unknown population values.

One other detail stands out in the graphs. There is much less variability in the estimates using stratified random sampling with the rows as strata. The table on page N/DS-5 shows the actual number of healthy sunflowers in each grid square. Notice that the squares within each row contain a similar number of healthy plants but there are big differences between rows. *When we can choose strata that are "similar within but different between," stratified random samples give more precise estimates than simple random samples of the same size.*

Why didn't using the columns as strata reduce the variability of the estimates in a similar way? Because the squares within each column have very different numbers of healthy plants.



Both simple random sampling and stratified random sampling are hard to use when populations are large and spread out over a wide area. In that situation, we'd

prefer a method that selects groups (**clusters**) of individuals that are "near" one another. That's the idea of a **cluster sample.**

DEFINITION: Cluster sample and clusters

To get a **cluster sample**, start by classifying the population into groups of individuals that are located near each other, called **clusters**. Then choose an SRS of the clusters. All individuals in the chosen clusters are included in the sample.

In a cluster sample, some people take an SRS from each cluster rather than including all members of the cluster.

Cluster samples are often used for practical reasons, like saving time and money. Cluster sampling works best when the clusters look just like the population but on a smaller scale. Imagine a large high school that assigns its students to homerooms alphabetically by last name. The school administration is considering a new schedule and would like student input. Administrators decide to survey 200 randomly selected students. It would be difficult to track down an SRS of 200 students, so the administration opts for a cluster sample of homerooms. The principal (who knows some statistics) takes a simple random sample of 8 homerooms and gives the survey to all 25 students in each homeroom.

Cluster samples don't offer the statistical advantage of better information about the population that stratified random samples do. That's because clusters are often chosen for ease so they may have as much variability as the population itself.

Remember: strata are ideally "similar within, but different between," while clusters are ideally "different within, but similar between."

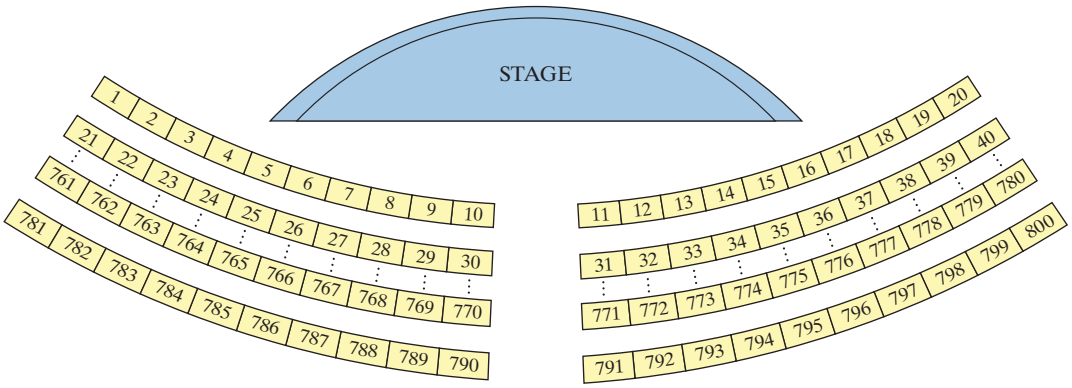
Be sure you understand the difference between strata and clusters. We want each stratum to contain similar individuals and for there to be large differences between strata. For a cluster sample, we'd *like* each cluster to look just like the population, but on a smaller scale. Here's an example that compares the random sampling methods we have discussed so far.

EXAMPLE Sampling at a School Assembly

Strata or clusters?

The student council wants to conduct a survey during the first five minutes of an all-school assembly in the auditorium about use of the school library. They would like to announce the results of the survey at the end of the assembly. The student council president asks your statistics class to help carry out the survey.

PROBLEM: There are 800 students present at the assembly. A map of the auditorium is shown on the next page. Note that students are seated by grade level and that the seats are numbered from 1 to 800.



9th grade: Seats 601–800	10th grade: Seats 401–600
11th grade: Seats 201–400	12th grade: Seats 1–200

Describe how you would use your calculator to select 80 students to complete the survey with each of the following:

(a) Simple random sample

(b) Stratified random sample

(c) Cluster sample

SOLUTION:

(a) To take an SRS, we need to choose 80 of the seat numbers at random. Use randInt(1,800) on your calculator until 80 different seats are selected. Then give the survey to the students in those seats.

(b) The students in the assembly are seated by grade level. Because students' library use might be similar within grade levels but different across grade levels, we'll use the grade level seating areas as our strata. Within each grade's seating area, we'll select 20 seats at random. For the 9th grade, use randInt(601,800) to select 20 different seats. Use randInt(401,600) to pick 20 different sophomore seats, randInt(201,400) to get 20 different junior seats, and randInt(1,200) to choose 20 different senior seats. Give the survey to the students in the selected seats.

(c) With the way students are seated, each column of seats from the stage to the back of the auditorium could be used as a cluster. Note that each cluster contains students from all four grade levels, so each should represent the population well. Because there are 20 clusters, each with 40 seats, we need to choose 2 clusters at random to get 80 students for the survey. Use randInt(1,20) to select two clusters, and then give the survey to all 40 students in each column of seats.

Note that cluster sampling is much more efficient than finding 80 seats scattered about the auditorium, as required by both of the other sampling methods.

For Practice *Try Exercise* 21

Most large-scale sample surveys use *multistage samples* that combine two or more sampling methods. For example, the U.S. Census Bureau carries out a monthly Current Population Survey (CPS) of about 60,000 households. Researchers start by choosing a stratified random sample of neighborhoods in 756 of the 2007 geographical areas in the United States. Then they divide each neighborhood into clusters of four nearby households and select a cluster sample to interview.

Analyzing data from sampling methods more complex than an SRS takes us beyond basic statistics. But the SRS is the building block of more elaborate methods, and the principles of analysis remain much the same for these other methods.

CHECK YOUR UNDERSTANDING

The manager of a sports arena wants to learn more about the financial status of the people who are attending an NBA basketball game. He would like to give a survey to a representative sample of the more than 20,000 fans in attendance. Ticket prices for the game vary a great deal: seats near the court cost over $100 each, while seats in the top rows of the arena cost $25 each. The arena is divided into 30 numbered sections, from 101 to 130. Each section has rows of seats labeled with letters from A (nearest the court) to ZZ (top row of the arena).

1. Explain why it might be difficult to give the survey to an SRS of 200 fans.

2. Which would be a better way to take a stratified random sample of fans: using the lettered rows or the numbered sections as strata? Explain.

3. Which would be a better way to take a cluster sample of fans: using the lettered rows or the numbered sections as clusters? Explain.

Inference for Sampling

The purpose of a sample is to give us information about a larger population. The process of drawing conclusions about a population on the basis of sample data is called **inference** because we *infer* information about the population from what we *know* about the sample.

Inference from convenience samples or voluntary response samples would be misleading because these methods of choosing a sample are biased. We are almost certain that the sample does *not* fairly represent the population. *The first reason to rely on random sampling is to avoid bias in choosing a sample.*

Still, it is unlikely that results from a random sample are exactly the same as for the entire population. Sample results, like the unemployment rate obtained from the monthly Current Population Survey, are only estimates of the truth about the population. If we select two samples at random from the same population, we will almost certainly choose different individuals. So the sample results will differ somewhat, just by chance. Properly designed samples avoid systematic bias. But their results are rarely exactly correct, and we expect them to vary from sample to sample.

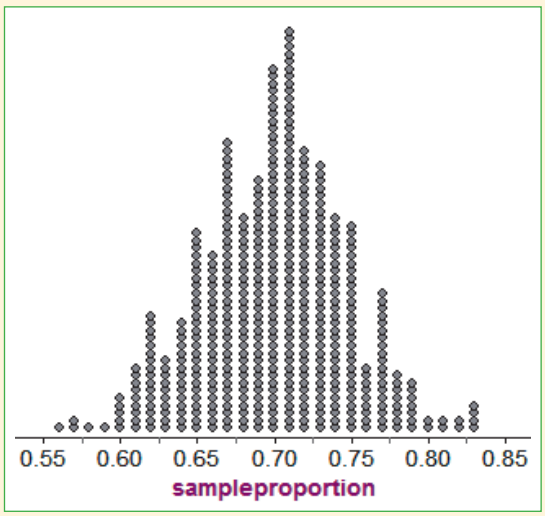
EXAMPLE

Going to class

How much do sample results vary?

Suppose that 70% of the students in a large university attended all their classes last week. Imagine taking a simple random sample of 100 students and recording the proportion of students in the sample who went to every class last week. Would the sample proportion be *exactly* 0.70? Probably not. Would the sample proportion be *close* to 0.70? That depends on what we mean by "close." The following graph shows the results of taking 500 SRSs, each of size 100, and recording the proportion of students who attended all their classes in each sample.

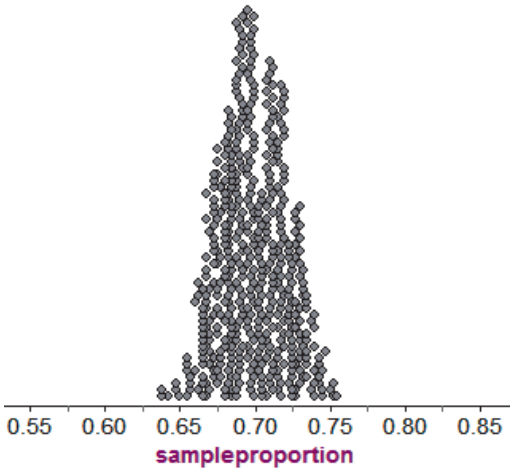
What do we see? The graph is centered at about 0.70, the population proportion. All of the sample proportions fall between 0.55 and 0.85. So we shouldn't be surprised if the difference between the sample proportion and the population proportion is as large as 0.15. The graph also has a very distinctive "bell shape."



Dotplot of the sample proportion of students in each of 500 SRSs of size 100 who attended all their classes last week. The population proportion is 0.70.

Why can we trust random samples? As the previous example illustrates, the results of random sampling don't change haphazardly from sample to sample. Because we deliberately use chance, the results obey the laws of probability that govern chance behavior. These laws allow us to say how likely it is that sample results are close to the truth about the population. The second reason to use random sampling is that the laws of probability allow trustworthy inference about the population. Results from random samples come with a "margin of error" that sets bounds on the size of the likely error. We will discuss the details of inference for sampling later.

One point is worth making now: *larger random samples give better information about the population than smaller samples.* For instance, let's look at what happens if we increase the sample size in the example from 100 to 400 students. The dotplot below shows the results of taking 500 SRSs, each of size 400, and recording the proportion of students who attended all their classes in each sample. This graph is also centered at about 0.70. But now all the sample proportions fall between 0.63 and 0.77. So the difference between the sample proportion and the population proportion is at most 0.07. When using SRSs of size 100, this difference could be as much as 0.15. The moral of the story: by taking a very large random sample, you can be confident that the sample result is very close to the truth about the population.



Dotplot of the sample proportion of students in each of 500 SRSs of size 400 who attended all their classes last week. The population proportion is 0.70.

The Current Population Survey contacts about 60,000 households, so we'd expect its estimate of the national unemployment rate to be within about 0.1% of the actual population value. Opinion polls that contact 1000 or 1500 people give less precise results—we expect the sample result to be within about 3% of the actual population percent with a given opinion. Of course, only samples chosen by chance carry this guarantee. Lou Dobbs's online sample tells us little about overall American public opinion even though 7350 people clicked a response.

Sample Surveys: What Can Go Wrong?

The use of bad sampling methods (convenience or voluntary response) often leads to bias. Researchers can avoid bad methods by using random sampling to choose their samples. Other problems in conducting sample surveys are more difficult to avoid.

The list of individuals from which a sample will be drawn is called the *sampling frame.*

Sampling is often done using a list of individuals in the population. Such lists are seldom accurate or complete. The result is **undercoverage.**

DEFINITION: Undercoverage

Undercoverage occurs when some members of the population cannot be chosen in a sample.

Most samples suffer from some degree of undercoverage. A sample survey of households, for example, will miss not only homeless people but also prison inmates and students in dormitories. An opinion poll conducted by calling landline telephone numbers will miss households that have only cell phones as well as households without a phone. The results of national sample surveys therefore have some bias due to undercoverage if the people not covered differ from the rest of the population.

Well-designed sample surveys avoid bias in the sampling process. The real problems start *after* the sample is chosen.

One of the most serious sources of bias in sample surveys is **nonresponse**.

DEFINITION: Nonresponse

Nonresponse occurs when an individual chosen for the sample can't be contacted or refuses to participate.

Nonresponse to surveys often exceeds 50%, even with careful planning and several follow-up calls. If the people who respond differ from those who don't, in a way that is related to the response, bias results.

Some students misuse the term "voluntary response" to explain why certain individuals don't respond in a sample survey. Their idea is that participation in the survey is optional (voluntary), so anyone can refuse to take part. What the students are describing is nonresponse. Think about it this way: nonresponse can occur only after a sample has been selected. In a voluntary response sample, every individual has opted to take part, so there won't be any nonresponse.

EXAMPLE

The ACS, GSS, and Opinion Polls

How bad is nonresponse?

The Census Bureau's American Community Survey (ACS) has the lowest nonresponse rate of any poll we know: only about 1% of the households in the sample refuse to respond. The overall nonresponse rate, including "never at home" and other causes, is just 2.5%.[10] This monthly survey of about 250,000 households replaces the "long form" that in the past was sent to some households in the every-ten-years national census. Participation in the ACS is mandatory, and the Census Bureau follows up by telephone and then in person if a household doesn't return the mail questionnaire.

The University of Chicago's General Social Survey (GSS) is the nation's most important social science survey (see Figure 4.1). The GSS contacts its sample in person, and it is run by a university. Despite these advantages, its most recent survey had a 30% rate of nonresponse.

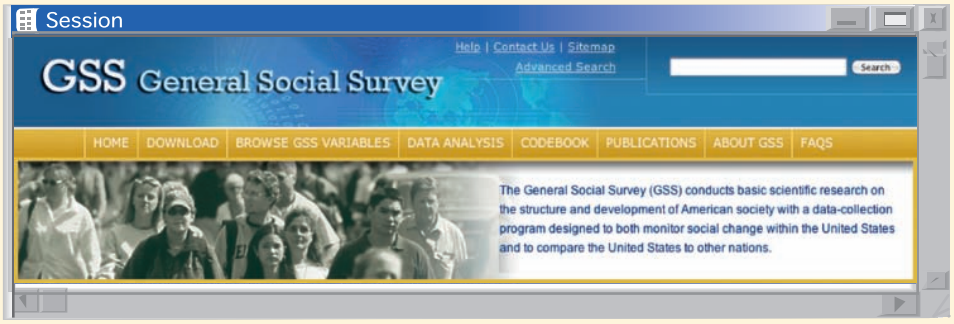
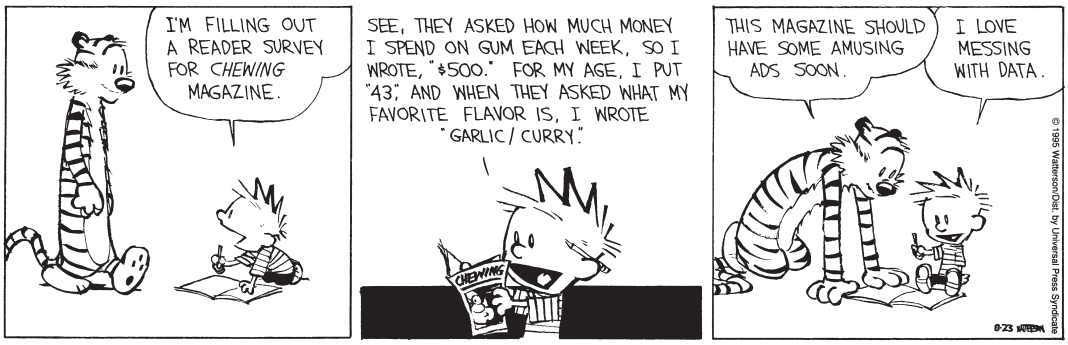


FIGURE 4.1 The home page of the General Social Survey at the University of Chicago's National Opinion Research Center (http://www3.norc.org/GSS+Website/). The GSS has tracked opinions about a wide variety of issues since 1972.

What about opinion polls by news media and opinion-polling firms? We don't know their rates of nonresponse because they won't say. That's a bad sign. The Pew Research Center for the People and the Press imitated a careful random digit dialing survey and published the results: over 5 days, the survey reached 76% of the households in its chosen sample, but "because of busy schedules, skepticism and outright refusals, interviews were completed in just 38% of households that were reached." Combining households that could not be contacted with those who did not complete the interview gave a nonresponse rate of 73%.[11]

Another type of nonsampling problem occurs when people give inaccurate answers to survey questions. People may lie about their age, income, or drug use. They may misremember how many hours they spent on the Internet last week. Or they might make up an answer to a question that they don't understand.

The gender, race, age, ethnicity, or behavior of the interviewer can also affect people's responses. A systematic pattern of inaccurate answers in a survey leads to **response bias**.



The **wording of questions** is the most important influence on the answers given to a sample survey. Confusing or leading questions can introduce strong bias. Changes in wording can greatly affect a survey's outcome.

EXAMPLE

How Do Americans Feel about Illegal Immigrants?

Question wording matters

"Should illegal immigrants be prosecuted and deported for being in the U.S. illegally, or shouldn't they?" Asked this question in an opinion poll, 69% favored deportation. But when the very same sample was asked whether illegal immigrants who have worked in the United States for two years "should be given a chance to keep their jobs and eventually apply for legal status," 62% said that they should. Different questions give quite different impressions of attitudes toward illegal immigrants.

Even the order in which questions are asked matters. *Don't trust the results of a sample survey until you have read the exact questions asked.*

caution

THINK ABOUT IT

Does the order matter? Ask a sample of college students these two questions:

"How happy are you with your life in general?" (Answers on a scale of 1 to 5)
"How many dates did you have last month?"

There is almost no association between responses to the two questions when asked in this order. It appears that dating has little to do with happiness. Reverse the order of the questions, however, and a much stronger association appears: college students who say they had more dates tend to give higher ratings of happiness about life. Asking a question that brings dating to mind makes dating success a big factor in happiness.

CHECK YOUR UNDERSTANDING

1. Each of the following is a possible source of bias in a sample survey. Name the type of bias that could result.

(a) The sample is chosen at random from a telephone directory.

(b) Some people cannot be contacted in five calls.

(c) Interviewers choose people walking by on the sidewalk to interview.

2. A survey paid for by makers of disposable diapers found that 84% of the sample opposed banning disposable diapers. Here is the actual question:

> It is estimated that disposable diapers account for less than 2% of the trash in today's landfills. In contrast, beverage containers, third-class mail, and yard wastes are estimated to account for about 21% of the trash in landfills. Given this, in your opinion, would it be fair to ban disposable diapers?[12]

Explain how the wording of the question could result in bias. Be sure to specify the direction of the bias.

Section 4.1 Summary

- A **census** collects data from every individual in the **population**.
- A **sample survey** selects a **sample** from the population of all individuals about which we desire information. The goal of a sample survey is **inference:** we draw conclusions about the population based on data from the sample. It is important to specify exactly what population you are interested in and what variables you will measure.
- **Convenience samples** choose individuals who are easiest to reach. In **voluntary response samples**, individuals choose to join the sample in response to an open invitation. Both of these sampling methods usually lead to **bias**: they consistently underestimate or consistently overestimate the value you want to know.
- **Random sampling** uses chance to select a sample.
- The basic random sampling method is a **simple random sample (SRS).** An SRS gives every possible sample of a given size the same chance to be chosen. Choose an SRS by labeling the members of the population and using slips of paper, random digits, or technology to select the sample.
- To choose a **stratified random sample,** divide the population into **strata,** groups of individuals that are similar in some way that might affect their responses. Then choose a separate SRS from each stratum and combine these SRSs to form the sample. When strata are "similar within but different between," stratified random samples tend to give more precise estimates of unknown population values than simple random samples.
- To choose a **cluster sample,** divide the population into groups of individuals that are located near each other, called **clusters.** Randomly select some of these clusters. All the individuals in the chosen clusters are included in the sample. Ideally, clusters are "different within but similar between." Cluster

sampling saves time and money by collecting data from entire groups of individuals that are close together.

- Random sampling helps avoid bias in choosing a sample. Bias can still occur in the sampling process due to **undercoverage**, which happens when some members of the population cannot be chosen.
- The most serious errors in sample surveys, however, are ones that occur after the sample is chosen. The single biggest problem is **nonresponse:** when people can't be contacted or refuse to answer. Incorrect answers by respondents can lead to **response bias**. Finally, the **wording of questions** has a big influence on the answers.

4.1 TECHNOLOGY CORNER

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

10. Choosing an SRS page 215

Section 4.1 Exercises

1. **Students as customers** A high school's student newspaper plans to survey local businesses about the importance of students as customers. From an alphabetical list of all local businesses, the newspaper staff chooses 150 businesses at random. Of these, 73 return the questionnaire mailed by the staff. Identify the population and the sample. pg 210

2. **Student archaeologists** An archaeological dig turns up large numbers of pottery shards, broken stone tools, and other artifacts. Students working on the project classify each artifact and assign it a number. The counts in different categories are important for understanding the site, so the project director chooses 2% of the artifacts at random and checks the students' work. Identify the population and the sample.

3. **Sampling stuffed envelopes** A large retailer prepares its customers' monthly credit card bills using an automatic machine that folds the bills, stuffs them into envelopes, and seals the envelopes for mailing. Are the envelopes completely sealed? Inspectors choose 40 envelopes at random from the 1000 stuffed each hour for visual inspection. Identify the population and the sample.

4. **Customer satisfaction** A department store mails a customer satisfaction survey to people who make credit card purchases at the store. This month, 45,000 people made credit card purchases. Surveys are mailed to 1000 of these people, chosen at random, and 137 people return the survey form. Identify the population and the sample.

5. **Call the shots** An advertisement for an upcoming TV show asked: "Should handgun control be tougher? You call the shots in a special call-in poll tonight. If yes, call 1-900-720-6181. If no, call 1-900-720-6182. Charge is 50 cents for the first minute." Over 90% of people who called in said "Yes." Explain why this opinion poll is almost certainly biased.

6. **Explain it to the congresswoman** You are on the staff of a member of Congress who is considering a bill that would provide government-sponsored insurance for nursing-home care. You report that 1128 letters have been received on the issue, of which 871 oppose the legislation. "I'm surprised that most of my constituents oppose the bill. I thought it would be quite popular," says the congresswoman. Are you convinced that a majority of the voters oppose the bill? How would you explain the statistical issue to the congresswoman?

7. **Instant opinion** A recent online poll posed the question "Should female athletes be paid the same as men for the work they do?" In all, 13,147 (44%) said "Yes," 15,182 (50%) said "No," and the

remaining 1448 said "Don't know." In spite of the large sample size for this survey, we can't trust the result. Why not?

8. **Sampling at the mall** You have probably seen the mall interviewer, approaching people passing by with clipboard in hand. Explain why even a large sample of mall shoppers would not provide a trustworthy estimate of the current unemployment rate.

9. **Sleepless nights** How much sleep do high school students get on a typical school night? An interested student designed a survey to find out. To make data collection easier, the student surveyed the first 100 students to arrive at school on a particular morning. These students reported an average of 7.2 hours of sleep on the previous night.

pg 213

(a) What type of sample did the student obtain?

(b) Explain why this sampling method is biased. Is 7.2 hours probably higher or lower than the true average amount of sleep last night for all students at the school? Why?

10. **Online polls** In June 2008, *Parade* magazine posed the following question: "Should drivers be banned from using all cell phones?" Readers were encouraged to vote online at parade.com. The July 13, 2008, issue of *Parade* reported the results: 2407 (85%) said "Yes" and 410 (15%) said "No."

(a) What type of sample did the *Parade* survey obtain?

(b) Explain why this sampling method is biased. Is 85% probably higher or lower than the true percent of all adults who believe that cell phone use while driving should be banned? Why?

11. **Do you trust the Internet?** You want to ask a sample of high school students the question "How much do you trust information about health that you find on the Internet—a great deal, somewhat, not much, or not at all?" You try out this and other questions on a pilot group of 5 students chosen from your class. The class members are listed below.

pg 217

(a) Explain how you would use a line of Table D to choose an SRS of 5 students from the following list. Explain your method clearly enough for a classmate to obtain your results.

(b) Use line 107 to select the sample. Show how you use each of the digits.

Anderson	Deng	Glaus	Nguyen	Samuels
Arroyo	De Ramos	Helling	Palmiero	Shen
Batista	Drasin	Husain	Percival	Tse
Bell	Eckstein	Johnson	Prince	Velasco
Burke	Fernandez	Kim	Puri	Wallace
Cabrera	Fullmer	Molina	Richards	Washburn
Calloway	Gandhi	Morgan	Rider	Zabidi
Delluci	Garcia	Murphy	Rodriguez	Zhao

12. **Apartment living** You are planning a report on apartment living in a college town. You decide to select three apartment complexes at random for in-depth interviews with residents.

(a) Explain how you would use a line of Table D to choose an SRS of 3 complexes from the list below. Explain your method clearly enough for a classmate to obtain your results.

(b) Use line 117 to select the sample. Show how you use each of the digits.

Ashley Oaks	Chauncey Village	Franklin Park	Richfield
Bay Pointe	Country Squire	Georgetown	Sagamore Ridge
Beau Jardin	Country View	Greenacres	Salem Courthouse
Bluffs	Country Villa	Lahr House	Village Manor
Brandon Place	Crestview	Mayfair Village	Waterford Court
Briarwood	Del-Lynn	Nobb Hill	Williamsburg
Brownstone	Fairington	Pemberly Courts	
Burberry	Fairway Knolls	Peppermill	
Cambridge	Fowler	Pheasant Run	

13. **Sampling the forest** To gather data on a 1200-acre pine forest in Louisiana, the U.S. Forest Service laid a grid of 1410 equally spaced circular plots over a map of the forest. A ground survey visited a sample of 10% of these plots.[13]

(a) Explain how you would use your calculator or Table D to choose an SRS of 141 plots. Your description should be clear enough for a classmate to carry out your plan.

(b) Use your method from (a) to choose the first 3 plots.

14. **Sampling gravestones** The local genealogical society in Coles County, Illinois, has compiled records on all 55,914 gravestones in cemeteries in the county for the years 1825 to 1985. Historians plan to use these records to learn about African Americans in Coles County's history. They first choose an SRS of 395 records to check their accuracy by visiting the actual gravestones.[14]

(a) Explain how you would use your calculator or Table D to choose the SRS. Your description should be clear enough for a classmate to carry out your plan.

(b) Use your method from (a) to choose the first 3 gravestones.

15. **Random digits** Which of the following statements are true of a table of random digits, and which are false? Briefly explain your answers.

(a) There are exactly four 0s in each row of 40 digits.

(b) Each pair of digits has chance 1/100 of being 00.

(c) The digits 0000 can never appear as a group, because this pattern is not random.

16. **Random digits** In using Table D repeatedly to choose random samples, you should not always begin at the same place, such as line 101. Why not?

17. **iPhones** Suppose 1000 iPhones are produced at a factory today. Management would like to ensure that the phones' display screens meet their quality standards before shipping them to retail stores. Since it takes about 10 minutes to inspect an individual phone's display screen, managers decide to inspect a sample of 20 phones from the day's production.

(a) Explain why it would be difficult for managers to inspect an SRS of 20 iPhones that are produced today.

(b) An eager employee suggests that it would be easy to inspect the last 20 iPhones that were produced today. Why isn't this a good idea?

(c) Another employee recommends a different sampling method: Randomly choose one of the first 50 iPhones produced. Inspect that phone and every fiftieth iPhone produced afterward. (This method is known as **systematic random sampling**.) Explain carefully why this sampling method is *not* an SRS.

18. **Dead trees** On the west side of Rocky Mountain National Park, many mature pine trees are dying due to infestation by pine beetles. Scientists would like to use sampling to estimate the proportion of all pine trees in the area that have been infected.

(a) Explain why it wouldn't be practical for scientists to obtain an SRS in this setting.

(b) A possible alternative would be to use every pine tree along the park's main road as a sample. Why is this sampling method biased?

(c) Suppose that a more complicated random sampling plan is carried out, and that 35% of the pine trees in the sample are infested by the pine beetle. Can scientists conclude that exactly 35% of *all* the pine trees on the west side of the park are infested? Why or why not?

19. **Who goes to the convention?** A club has 30 student members and 10 faculty members. The students are

Abel	Fisher	Huber	Miranda	Reinmann
Carson	Ghosh	Jimenez	Moskowitz	Santos
Chen	Griswold	Jones	Neyman	Shaw
David	Hein	Kim	O'Brien	Thompson
Deming	Hernandez	Klotz	Pearl	Utts
Elashoff	Holland	Liu	Potter	Varga

The faculty members are

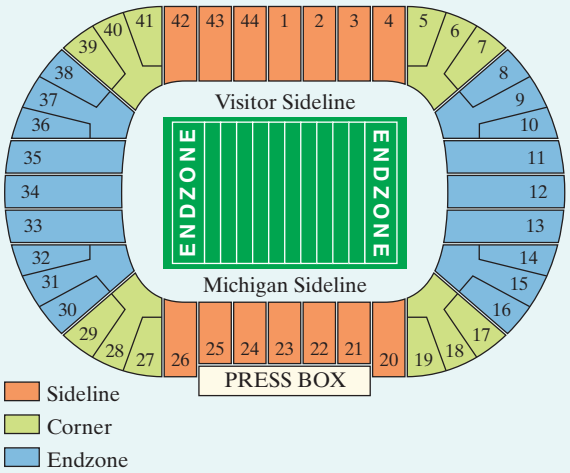
Andrews	Fernandez	Kim	Moore	West
Besicovitch	Gupta	Lightman	Phillips	Yang

The club can send 4 students and 2 faculty members to a convention. It decides to choose those who will go by random selection. Describe a method for using Table D to select a stratified random sample of 4 students and 2 faculty. Then use line 123 to select the sample.

20. **Sampling by accountants** Accountants often use stratified samples during audits to verify a company's records of such things as accounts receivable. The stratification is based on the dollar amount of the item and often includes 100% sampling of the largest items. One company reports 5000 accounts receivable. Of these, 100 are in amounts over \$50,000; 500 are in amounts between \$1000 and \$50,000; and the remaining 4400 are in amounts under \$1000. Using these groups as strata, you decide to verify all the largest accounts and to sample 5% of the midsize accounts and 1% of the small accounts. Describe a method for using Table D to select a stratified random sample of the midsize and small accounts. Then use line 115 to select only the first 3 accounts from each of these strata.

21. **Go Blue!** Michigan Stadium, also known as "The Big House," seats over 100,000 fans for a football game. The University of Michigan athletic department plans to conduct a survey about concessions that are sold during games. Tickets are most expensive for seats on the sidelines. The cheapest seats are in the end zones (where one of the authors sat as a student). A map of the stadium is shown.

pg 221



(a) The athletic department is considering a stratified random sample. What would you recommend as the strata? Why?

(b) Explain why a cluster sample might be easier to obtain. What would you recommend for the clusters? Why?

22. **How was your stay?** A hotel has 30 floors with 40 rooms per floor. The rooms on one side of the hotel face the water, while rooms on the other side face a golf course. There is an extra charge for the rooms with a water view. The hotel manager wants to survey 120 guests who stayed at the hotel during a convention about their overall satisfaction with the property.

(a) Explain why choosing a stratified random sample might be preferable to an SRS in this case. What would you use as strata?

(b) Why might a cluster sample be a simpler option? What would you use as clusters?

23. **Is it an SRS?** A corporation employs 2000 male and 500 female engineers. A stratified random sample of 200 male and 50 female engineers gives each engineer 1 chance in 10 to be chosen. This sample design gives every individual in the population the same chance to be chosen for the sample. Is it an SRS? Explain your answer.

24. **Attitudes toward alcohol** At a party there are 30 students over age 21 and 20 students under age 21. You choose at random 3 of those over 21 and separately choose at random 2 of those under 21 to interview about attitudes toward alcohol. You have given every student at the party the same chance to be interviewed. Why is your sample not an SRS?

25. **High-speed Internet** Laying fiber-optic cable is expensive. Cable companies want to make sure that if they extend their lines out to less dense suburban or rural areas, there will be sufficient demand and the work will be cost-effective. They decide to conduct a survey to determine the proportion of households in a rural subdivision that would buy the service. They select a simple random sample of 5 blocks in the subdivision and survey each family that lives on one of those blocks.

(a) What is the name for this kind of sampling method?

(b) Give a possible reason why the cable company chose this method.

26. **Timber!** A lumber company wants to estimate the proportion of trees in a large forest that are ready to be cut down. They use an aerial map to divide the forest into 200 equal-sized rectangles. Then they choose a random sample of 20 rectangles and examine every tree that's in one of those rectangles.

(a) What is the name for this kind of sampling method?

(b) Give a possible reason why the lumber company chose this method.

27. **Tweet, tweet!** What proportion of students at your school use Twitter? To find out, you survey a simple random sample of students from the school roster.

(a) Will your sample result be exactly the same as the true population proportion? Explain.

(b) Which would be more likely to get your sample result closer to the true population value: an SRS of 50 students or an SRS of 100 students? Explain.

28. **Far from home?** A researcher wants to estimate the average distance that students at a large community college live from campus. To find out, she surveys a simple random sample of students from the registrar's database.

(a) Will the researcher's sample result be exactly the same as the true population mean? Explain.

(b) Which would be more likely to get the researcher's sample result closer to the true population value: an SRS of 100 students or an SRS of 50 students? Explain.

29. **Baseball tickets** Suppose you want to know the average amount of money spent by the fans attending opening day for the Cleveland Indians baseball season. You get permission from the team's management to conduct a survey at the stadium, but they will not allow you to bother the fans in the club seating or box seats (the most expensive seating). Using a computer, you randomly select 500 seats from the rest of the stadium. During the game, you ask the fans in those seats how much they spent that day.

Give a reason why this survey might yield a biased result. Explain the likely direction of the bias.

30. **Rise and shine** How long before school starts do students get out of bed, on average? Administrators survey a random sample of students on each school bus one morning.

Give a reason why this survey might yield a biased result. Explain the likely direction of the bias.

31. **Nonresponse** A survey of drivers began by randomly sampling all listed residential telephone numbers in the United States. Of 45,956 calls to these numbers, 5029 were completed. The goal of the survey was to estimate how far people drive, on average, per day.[15]

(a) What was the rate of nonresponse for this sample?

(b) Explain how nonresponse can lead to bias in this survey. Be sure to give the direction of the bias.

32. **Ring-no-answer** A common form of nonresponse in telephone surveys is "ring-no-answer." That is, a call is made to an active number but no one answers. The Italian National Statistical Institute looked at nonresponse to a government survey of households in Italy during the periods January 1 to Easter and July 1 to August 31. All calls were made between 7 and 10 P.M., but 21.4% gave "ring-no-answer" in one period versus 41.5% "ring-no-answer" in the other period.[16] Which period do you think had the higher rate of no answers? Why? Explain why a high rate of nonresponse makes sample results less reliable.

33. **Running red lights** The sample described in Exercise 31 produced a list of 5024 licensed drivers. The investigators then chose an SRS of 880 of these drivers to answer questions about their driving habits. One question asked was: "Recalling the last ten traffic lights you drove through, how many of them were red when you entered the intersections?" Of the 880 respondents, 171 admitted that at least one light had been red. A practical problem with this survey is that

people may not give truthful answers. What is the likely direction of the bias? Explain.

34. **Seat belt use** A study in El Paso, Texas, looked at seat belt use by drivers. Drivers were observed at randomly chosen convenience stores. After they left their cars, they were invited to answer questions that included questions about seat belt use. In all, 75% said they always used seat belts, yet only 61.5% were wearing seat belts when they pulled into the store parking lots.[17] Explain the reason for the bias observed in responses to the survey. Do you expect bias in the same direction in most surveys about seat belt use?

35. **Wording bias** Comment on each of the following as a potential sample survey question. Is the question clear? Is it slanted toward a desired response?

(a) "Some cell phone users have developed brain cancer. Should all cell phones come with a warning label explaining the danger of using cell phones?"

(b) "Do you agree that a national system of health insurance should be favored because it would provide health insurance for everyone and would reduce administrative costs?"

(c) "In view of escalating environmental degradation and incipient resource depletion, would you favor economic incentives for recycling of resource–intensive consumer goods?"

36. **Checking for bias** Comment on each of the following as a potential sample survey question. Is the question clear? Is it slanted toward a desired response?

(a) Which of the following best represents your opinion on gun control?

1. The government should confiscate our guns.
2. We have the right to keep and bear arms.

(b) A freeze in nuclear weapons should be favored because it would begin a much-needed process to stop everyone in the world from building nuclear weapons now and reduce the possibility of nuclear war in the future. Do you agree or disagree?

Multiple choice: Select the best answer for Exercises 37 to 42.

37. The Web portal AOL places opinion poll questions next to many of its news stories. Simply click your response to join the sample. One of the questions in January 2008 was "Do you plan to diet this year?" More than 30,000 people responded, with 68% saying "Yes." You can conclude that

(a) about 68% of Americans planned to diet in 2008.

(b) the poll used a convenience sample, so the results tell us little about the population of all adults.

(c) the poll uses voluntary response, so the results tell us little about the population of all adults.

(d) the sample is too small to draw any conclusion.

(e) None of these.

38. To gather information about the validity of a new standardized test for tenth-grade students in a particular state, a random sample of 15 high schools was selected from the state. The new test was administered to every 10th-grade student in the selected high schools. What kind of sample is this?

(a) A simple random sample

(b) A stratified random sample

(c) A cluster sample

(d) A systematic random sample

(e) A voluntary response sample

39. Your statistics class has 30 students. You want to call an SRS of 5 students from your class to ask where they use a computer for the online quizzes. You label the students 01, 02, . . . , 30. You enter the table of random digits at this line:

14459 26056 31424 80371 65103 62253 22490 61181

Your SRS contains the students labeled

(a) 14, 45, 92, 60, 56.

(b) 14, 31, 03, 10, 22.

(c) 14, 03, 10, 22, 22.

(d) 14, 03, 10, 22, 06.

(e) 14, 03, 10, 22, 11.

40. Suppose that 35% of the registered voters in a state are registered as Republicans, 40% as Democrats, and 25% as Independents. A newspaper wants to select a sample of 1000 registered voters to predict the outcome of the next election. If they randomly select 350 Republicans, randomly select 400 Democrats, and randomly select 250 Independents, did this sampling procedure result in a simple random sample of registered voters from this district?

(a) Yes, because each registered voter had the same chance of being chosen.

(b) Yes, because random chance was involved.

(c) No, because not all registered voters had the same chance of being chosen.

(d) No, because there were a different number of registered voters selected from each party.

(e) No, because not all possible groups of 1000 registered voters had the same chance of being chosen.

41. A local news agency conducted a survey about unemployment by randomly dialing phone numbers until they had gathered responses from 1000 adults in their state. In the survey, 19% of those who responded said they were not currently employed. In reality, only 6% of the adults in the state were not currently employed

at the time of the survey. Which of the following best explains the difference in the two percentages?

(a) The difference is due to sampling variability. We shouldn't expect the results of a random sample to match the truth about the population every time.

(b) The difference is due to response bias. Adults who are employed are likely to lie and say that they are unemployed.

(c) The difference is due to undercoverage bias. The survey included only adults and did not include teenagers who are eligible to work.

(d) The difference is due to nonresponse bias. Adults who are employed are less likely to be available for the sample than adults who are unemployed.

(e) The difference is due to voluntary response. Adults are able to volunteer as a member of the sample.

42. A simple random sample of 1200 adult Americans is selected, and each person is asked the following question: "In light of the huge national deficit, should the government at this time spend additional money to establish a national system of health insurance?" Only 39% of those responding answered "Yes." This survey

(a) is reasonably accurate since it used a large simple random sample.

(b) needs to be larger since only about 24 people were drawn from each state.

(c) probably understates the percent of people who favor a system of national health insurance.

(d) is very inaccurate but neither understates nor overstates the percent of people who favor a system of national health insurance. Because simple random sampling was used, it is unbiased.

(e) probably overstates the percent of people who favor a system of national health insurance.

43. **Sleep debt** (3.2) A researcher reported that the typical teenager needs 9.3 hours of sleep per night but gets only 6.3 hours.[18] By the end of a 5-day school week, a teenager would accumulate about 15 hours of "sleep debt." Students in a high school statistics class were skeptical, so they gathered data on the amount of sleep debt (in hours) accumulated over time (in days) by a random sample of 25 high school students. The resulting least-squares regression equation for their data is $\widehat{\text{Sleep debt}} = 2.23 + 3.17(\text{days})$.

(a) Interpret the slope of the regression line in context.

(b) Are the students' results consistent with the researcher's report? Explain.

44. **Internet charges** (2.1) Some Internet service providers (ISPs) charge companies based on how much bandwidth they use in a month. One method that ISPs use for calculating bandwidth is to find the 95th percentile of a company's usage based on samples of hundreds of 5-minute intervals during a month.

(a) Explain what "95th percentile" means in this setting.

(b) Which would cost a company more: the 95th percentile method or a similar approach using the 98th percentile? Justify your answer.

4.2 Experiments

WHAT YOU WILL LEARN **By the end of the section, you should be able to:**

- Distinguish between an observational study and an experiment.
- Explain the concept of confounding and how it limits the ability to make cause-and-effect conclusions.
- Identify the experimental units, explanatory and response variables, and treatments in an experiment.
- Explain the purpose of comparison, random assignment, control, and replication in an experiment.
- Describe a completely randomized design for an experiment, including how to randomly assign treatments using slips of paper, technology, or a table of random digits.
- Describe the placebo effect and the purpose of blinding in an experiment.
- Interpret the meaning of statistically significant in the context of an experiment.
- Explain the purpose of blocking in an experiment. Describe a randomized block design or a matched pairs design for an experiment.

A sample survey aims to gather information about a population without disturbing the population in the process. Sample surveys are one kind of **observational study.** Other observational studies watch the behavior of animals in the wild or the interactions between teacher and students in the classroom. This section is about statistical designs for **experiments,** a very different way to produce data.

Observational Study versus Experiment

In contrast to observational studies, experiments don't just observe individuals or ask them questions. They actively impose some *treatment* to measure the response. Experiments can answer questions like "Does aspirin reduce the chance of a heart attack?" and "Can yoga help dogs live longer?"

DEFINITION: Observational study and experiment

An **observational study** observes individuals and measures variables of interest but does not attempt to influence the responses.

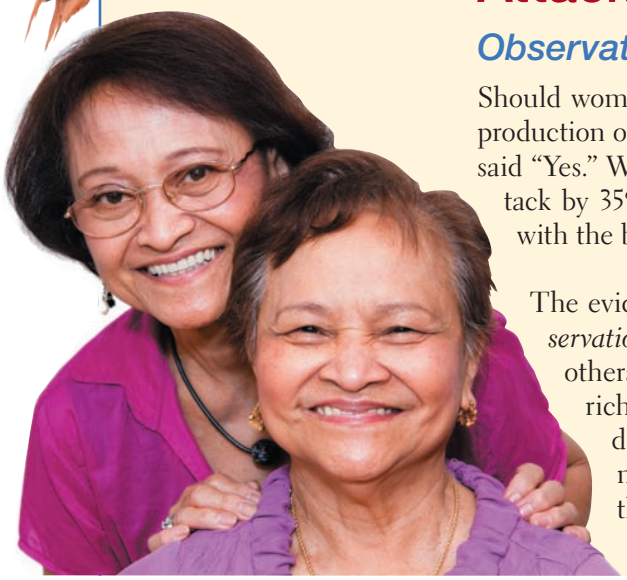
An **experiment** deliberately imposes some treatment on individuals to measure their responses.

The goal of an observational study can be to describe some group or situation, to compare groups, or to examine relationships between variables. The purpose of an experiment is to determine whether the treatment causes a change in the response. An observational study, even one based on a random sample, is a poor way to gauge the effect that changes in one variable have on another variable. To see the response to a change, we must actually impose the change. *When our goal is to understand cause and effect, experiments are the only source of fully convincing data.* For this reason, the distinction between observational study and experiment is one of the most important in statistics.

EXAMPLE

Does Taking Hormones Reduce Heart Attack Risk after Menopause?

Observation versus experiment



Should women take hormones such as estrogen after menopause, when natural production of these hormones ends? In 1992, several major medical organizations said "Yes." Women who took hormones seemed to reduce their risk of a heart attack by 35% to 50%. The risks of taking hormones appeared small compared with the benefits.

The evidence in favor of hormone replacement came from a number of *observational studies* that compared women who were taking hormones with others who were not. But the women who chose to take hormones were richer and better educated and saw doctors more often than women who didn't take hormones. Because the women who took hormones did many other things to better maintain their health, it isn't surprising that they had fewer heart attacks.

To get convincing data on the link between hormone replacement and heart attacks, we should do an *experiment*. Experiments don't let women decide what to do. They assign women to either hormone replacement pills or to placebo pills that look and taste the same as the hormone pills. The assignment is done by a coin toss, so that all kinds of women are equally likely to get either treatment. By 2002, several experiments with women of different ages agreed that hormone replacement does *not* reduce the risk of heart attacks. The National Institutes of Health, after reviewing the evidence, concluded that the first studies were wrong. Taking hormones after menopause quickly fell out of favor.[19]

From Chapter 3: A **response variable** measures an outcome of a study. An **explanatory variable** may help explain or predict changes in a response variable.

For each of these studies, the *explanatory variable* was whether or not a woman took hormones, and the *response variable* was whether or not the woman had a heart attack. Researchers wanted to argue that changes in the explanatory variable (hormone status) actually caused changes in the response variable (heart attack status). In the early observational studies, however, the effect of taking hormones was mixed up with the characteristics of women who chose to take them. These other variables make it hard to see the true relationship between the explanatory and response variables.

Let's consider two other variables from the observational studies of hormone replacement: number of doctor visits per year and age. The women who chose to take hormones visited their doctors more often than the women who didn't take hormones. Did the women in the hormone group have fewer heart attacks because they got better health care or because they took hormones? We can't be sure. A situation like this, in which the effects of two variables on a response variable cannot be separated from each other, is called **confounding.**

Some people call a variable that results in confounding, like the number of doctor visits per year in this case, a *confounding variable.*

What about age? Older women are at greater risk of having a heart attack than younger women. If the women who took hormones were generally younger than those who didn't, we'd have more confounding. That wasn't the case, however. There was no link between age and group membership (hormones or not) in the observational studies. If there is no difference between the groups with respect to the other variable, there can be no confounding.

DEFINITION: Confounding

Confounding occurs when two variables are associated in such a way that their effects on a response variable cannot be distinguished from each other.

AP® EXAM TIP If you are asked to identify a possible confounding variable in a given setting, you are expected to explain how the variable you choose (1) is associated with the explanatory variable and (2) affects the response variable.

Observational studies of the effect of an explanatory variable on a response variable often fail because of confounding between the explanatory variable and one or more other variables. Well-designed experiments take steps to prevent confounding. The later hormone therapy experiments avoided confounding by letting chance decide who took hormones and who didn't. That way, women who took better care of themselves were split about evenly between the two groups. So were older women and younger women. When these experiments found no reduction in heart attack risk for women taking hormones, researchers began to doubt the results of the earlier observational studies. The moral of the story is simple: *beware the influence of other variables!*

CHECK YOUR UNDERSTANDING

1. Does reducing screen brightness increase battery life in laptop computers? To find out, researchers obtained 30 new laptops of the same brand. They chose 15 of the computers at random and adjusted their screens to the brightest setting. The other 15 laptop screens were left at the default setting—moderate brightness. Researchers then measured how long each machine's battery lasted. Was this an observational study or an experiment? Justify your answer.

Questions 2 to 4 refer to the following setting. Does eating dinner with their families improve students' academic performance? According to an ABC News article, "Teenagers who eat with their families at least five times a week are more likely to get better grades in school."[20] This finding was based on a sample survey conducted by researchers at Columbia University.

2. Was this an observational study or an experiment? Justify your answer.
3. What are the explanatory and response variables?
4. Explain clearly why such a study cannot establish a cause-and-effect relationship. Suggest a variable that may be confounded with whether families eat dinner together.

The Language of Experiments

An experiment is a statistical study in which we actually do something (a **treatment**) to people, animals, or objects (the **experimental units**) to observe the response. Here is the basic vocabulary of experiments.

DEFINITION: Treatment, experimental units, subjects

A specific condition applied to the individuals in an experiment is called a **treatment**. If an experiment has several explanatory variables, a treatment is a combination of specific values of these variables.

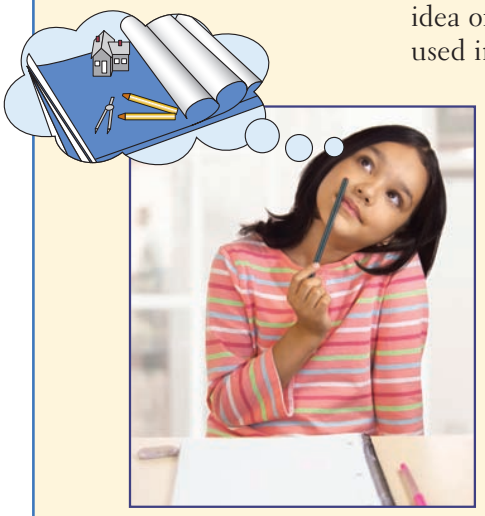
The **experimental units** are the smallest collection of individuals to which treatments are applied. When the units are human beings, they often are called **subjects**.

The best way to learn the language of experiments is to practice using it.

EXAMPLE When Will I Ever Use This Stuff?

Vocabulary of experiments

Researchers at the University of North Carolina were concerned about the increasing dropout rate in the state's high schools, especially for low-income students. Surveys of recent dropouts revealed that many of these students had started to lose interest during middle school. They said they saw little connection between what they were studying in school and their future plans. To change



this perception, researchers developed a program called CareerStart. The big idea of the program is that teachers show students how the topics they learn get used in specific careers.

To test the effectiveness of CareerStart, the researchers recruited 14 middle schools in Forsyth County to participate in an experiment. Seven of the schools, determined at random, used CareerStart along with the district's standard curriculum. The other seven schools just followed the standard curriculum. Researchers followed both groups of students for several years, collecting data on students' attendance, behavior, standardized test scores, level of engagement in school, and whether or not the students graduated from high school.

Results: Students at schools that used CareerStart generally had better attendance and fewer discipline problems, earned higher test scores, reported greater engagement in their classes, and were more likely to graduate.[21]

PROBLEM: Identify the experimental units, explanatory and response variables, and the treatments in the CareerStart experiment.

SOLUTION: The experimental units are 14 middle schools in Forsyth County, NC. The explanatory variable is whether the school used the CareerStart program with its students. Several response variables were measured, including test scores, attendance, behavior, student engagement, and graduation rates. This experiment compares two treatments: (1) the standard middle school curriculum and (2) the standard curriculum plus CareerStart.

For Practice *Try Exercise* 51

Note that the experimental units in the CareerStart example are the schools, not individual students. Experimental units are the smallest collection of individuals to which treatments are applied. The curricular treatments were administered to entire schools, so those are the experimental units.

The previous example illustrates the big advantage of experiments over observational studies: *experiments can give good evidence for causation*. In an experiment, we study the effects of the specific treatments we are interested in, while trying to control for the effects of other variables. For instance, the students in all 14 schools followed the standard curriculum. To ensure that the two groups of schools were as similar as possible before the treatments were administered, researchers let chance decide which 7 schools would use CareerStart. The only systematic difference between the schools was the educational treatment. When students from the CareerStart schools did much better, researchers were able to conclude that the program made the difference.

Sometimes, the explanatory variables in an experiment are called **factors**. Many experiments study the joint effects of several factors. In such an experiment, each treatment is formed by combining a specific value (often called a **level**) of each of the factors. Here's an example of a multifactor experiment.

EXAMPLE TV Advertising

Experiments with multiple explanatory variables

What are the effects of repeated exposure to an advertising message? The answer may depend on both the length of the ad and on how often it is repeated. An experiment investigated this question using 120 undergraduate students who volunteered to participate. All subjects viewed a 40-minute television program that included ads for a digital camera. Some subjects saw a 30-second commercial; others, a 90-second version. The same commercial was shown either 1, 3, or 5 times during the program. After viewing, all the subjects answered questions about their recall of the ad, their attitude toward the camera, and their intention to purchase it.[22]

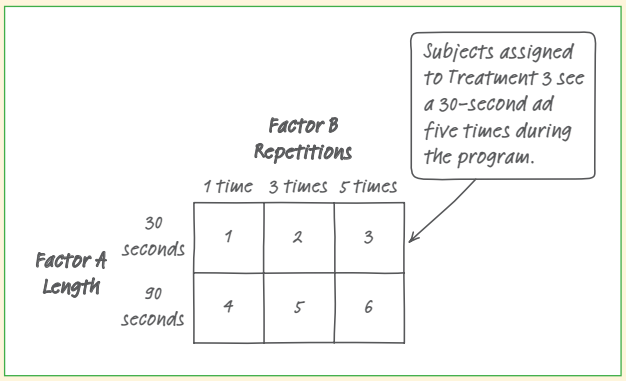


FIGURE 4.2 The six treatments in the TV ad experiment. Combinations of values of the two explanatory variables (factors) form six treatments.

PROBLEM: For the advertising study, identify the experimental units or subjects, explanatory and response variables, and the treatments.

SOLUTION:

The subjects are the 120 undergraduate students. This experiment has 2 explanatory variables (factors): length of the commercial and number of repetitions. The response variables include measures of subjects' recall of the ad, their attitudes about the digital camera, and whether they intend to purchase it.

There are 2 different lengths of commercial (30 and 90 seconds) and three different numbers of repetitions (1, 3, and 5). The 6 combinations consisting of one level of each factor form the 6 treatments shown in Figure 4.2: (1) 30 seconds, 1 time; (2) 30 seconds, 3 times; (3) 30 seconds, 5 times; (4) 90 seconds, 1 time; (5) 90 seconds, 3 times; (6) 90 seconds, 5 times.

For Practice *Try Exercise* 55

This example shows how experiments allow us to study the combined effect of several factors. The interaction of several factors can produce effects that could not be predicted from looking at the effect of each factor alone. Perhaps longer commercials increase interest in a product, and more commercials also increase interest. But if we both make a commercial longer and show it more often, viewers get annoyed and their interest in the product drops. The two-factor experiment in the TV advertising example will help us find out.

How to Experiment Badly

Experiments are the preferred method for examining the effect of one variable on another. By imposing the specific treatment of interest and controlling other influences, we can pin down cause and effect. Good designs are essential for effective experiments, just as they are for sampling. To see why, let's start with an example of a bad experimental design.

EXAMPLE

Are Online SAT Prep Courses Effective?

A bad experiment

A high school regularly offers a review course to prepare students for the SAT. This year, budget cuts will allow the school to offer only an online version of the course. Suppose the group of students who take the online course earn an average increase of 45 points in their math scores from a pre-test to the actual SAT test. Can we conclude that the online course is effective?

This experiment has a very simple design. A group of subjects (the students) were exposed to a treatment (the online course), and the outcome (increase in math scores) was observed. Here is the design:

Students → Online course → increase in math scores

A closer look showed that many of the students in the online review course were taking advanced math classes in school. Maybe the students in the online course improved their math scores because of what they were learning in their school math classes, not because of the online course. This confounding prevents us from concluding that the online course is effective.

Many laboratory experiments use a design like the one in the online SAT course example:

Experimental units → Treatment → Measure response

In the lab environment, simple designs often work well. Field experiments and experiments with animals or people deal with more varied conditions. *Outside the lab, badly designed experiments often yield worthless results because of confounding.*

How to Experiment Well

The remedy for the confounding in the SAT prep course example is to do a *comparative experiment* in which some students are taught the SAT course in the classroom and other, similar students take the course online. Most well-designed experiments compare two or more treatments.

Comparison alone isn't enough to produce results we can trust. If the treatments are given to groups that differ greatly when the experiment begins, *bias* will result. For example, if we allow students to select online or classroom instruction, more self-motivated students are likely to sign up for the online course. Allowing

personal choice will bias our results in the same way that volunteers bias the results of online opinion polls. The solution to the problem of bias in sampling is random selection. In experiments, the solution is **random assignment.**

DEFINITION: Random assignment

In an experiment, **random assignment** means that experimental units are assigned to treatments using a chance process.

Let's look at how random assignment can be used to improve the design of the SAT prep course experiment.

EXAMPLE SAT Prep: Online versus Classroom

How random assignment works

This year, the high school has enough budget money to compare the online SAT course with the classroom SAT course. Fifty students have agreed to participate in an experiment comparing the two instructional methods.

PROBLEM: Describe how you would randomly assign 25 students to each of the two methods:

(a) Using 50 identical slips of paper

(b) Using technology

(c) Using Table D

SOLUTION:

(a) The simplest way would be to use the "hat method." Write each subject's name on one of the slips. Put all the slips in a hat and mix them thoroughly. Draw them out one at a time until you have 25 slips. These 25 students will take the online course. The remaining 25 students will take the classroom course. Alternatively, you could write "online" on 25 of the slips and "classroom" on the other 25 slips. Then put the slips in a hat and mix them well. Have students come up one by one and (without looking) pick a slip from the hat. This guarantees 25 students per group, with the treatments assigned by chance.

(b) Give numbers 1, 2, 3, . . . , 49, 50 to the subjects in alphabetical order by last name. Then use your calculator's randInt command or a computer's random number generator to produce numbers between 1 and 50. Ignore any repeated numbers. The first 25 different numbers chosen select the students for the online course. The remaining 25 subjects will take the classroom course.

(c) Give labels 01, 02, 03, . . . , 49, 50 to the subjects in alphabetical order by last name. Go to a line of Table D and read two-digit groups moving from left to right. The first 25 distinct labels between 01 and 50 identify the 25 students that are assigned to the online course. The remaining 25 students will take the classroom course. Ignore repeated labels and groups of digits from 51 to 00.

For Practice *Try Exercise* 59

Random assignment should distribute the students taking advanced math classes in roughly equal numbers to each group. It should also balance out the number of students with lots of extracurricular activities and those with part-time jobs in the classroom and online SAT prep courses. Random assignment helps ensure that the effects of other variables (such as current math course or amount of available study time) are spread evenly among the two groups.

Although random assignment should create two groups of students that are roughly equivalent to begin with, we still have to ensure that the only consistent difference between the groups during the experiment is the type of SAT prep they receive. We can **control** for the effects of some variables by keeping them the same for both groups. For instance, we should give all students the same pretest and actual SAT test at the same times on the same days. The length, timing, content, and instructor of the SAT prep classes should also be the same.

Because the two groups are alike except for the treatments, any difference in their average math score improvements must be due either to the treatments themselves or to the random assignment. We can't say that any difference between the average SAT scores of students enrolled online and in the classroom must be caused by a difference in the effectiveness of the two types of instruction. There would be *some* difference, even if both groups received the same instruction, because of variation among students in background and study habits. Chance assigns students to one group or the other, which results in a chance difference between the groups.

In statistics, replication means "use enough subjects." In other fields, the term "replication" has a different meaning. When one experiment is conducted and then the same or a similar experiment is independently conducted in a different location by different investigators, this is known as replication. That is, replication means repeatability.

We would not trust an experiment with just one student in each group. The results would depend too much on which group got lucky and received the stronger student. If we assign many subjects to each group, however, the effects of chance will balance out, and there will be little difference in the average responses in the two groups unless the treatments themselves cause a difference. This is the idea of **replication:** use enough experimental units to distinguish a difference in the effects of the treatments from chance variation due to the random assignment.

PRINCIPLES OF EXPERIMENTAL DESIGN

The basic principles for designing experiments are as follows:

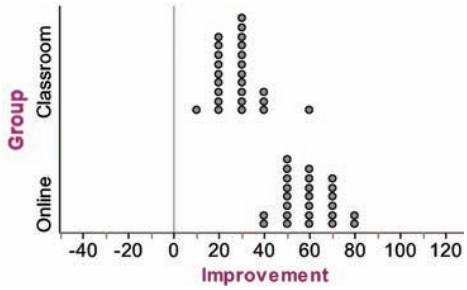
1. **Comparison.** Use a design that compares two or more treatments.
2. **Random assignment.** Use chance to assign experimental units to treatments. Doing so helps create roughly equivalent groups of experimental units by balancing the effects of other variables among the treatment groups.
3. **Control.** Keep other variables that might affect the response the same for all groups.
4. **Replication.** Use enough experimental units in each group so that any differences in the effects of the treatments can be distinguished from chance differences between the groups.

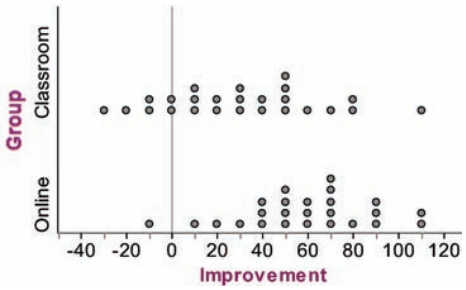
THINK ABOUT IT

Why is control important in an experiment? For two reasons. Suppose we used two different instructors in the SAT experiment. If Mrs. McDonald taught the online group and Mr. Tyson taught the classroom group, then course type will be confounded with instructor. We won't know if the difference in average improvement for the two groups was due to the difference in instructor or the difference in course type. So one reason we need to control other variables is to prevent confounding.

The second reason we should control other variables is to reduce variability in the response variable. Suppose that we allow students in both groups to choose how many class sessions to attend. Their choices will increase the variation in the response variable (improvement) for both groups. Some students will attend fewer sessions and experience smaller improvements than they would have otherwise. Other students will attend as many sessions as possible and experience bigger improvements than they might have otherwise. This increase in variation will make it harder to see if one treatment is really more effective.

The dotplots on the left show the results of an experiment in which the number of class sessions was the same for all participating students. From these graphs, it seems clear that the online course is more effective than the classroom course. The dotplots on the right show the results of an experiment in which the students were permitted to choose the number of class sessions they attended. Notice that the centers of the distributions haven't changed, but the distributions are much more variable. The increased overlap in the graphs makes the evidence supporting the online course less convincing.





Let's see how these principles were used in designing a famous medical experiment.

EXAMPLE

The Physicians' Health Study

A well-designed experiment

Does regularly taking aspirin help protect people against heart attacks? The Physicians' Health Study was a medical experiment that helped answer this question. In fact, the Physicians' Health Study looked at the effects of two drugs: aspirin and beta-carotene. Researchers wondered whether beta-carotene would help prevent some forms of cancer. The subjects in this experiment were 21,996 male physicians. There were two explanatory variables (factors), each having two levels: aspirin (yes or no) and beta-carotene (yes or no). Combinations of the levels of these factors form the four treatments shown in Figure 4.3 on the next page. One-fourth of the subjects were assigned at random to each of these treatments.

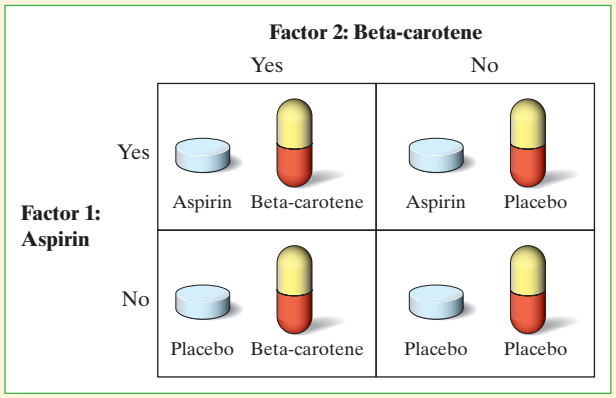


FIGURE 4.3 The treatments in the Physicians' Health Study.

On odd-numbered days, the subjects took either a tablet that contained aspirin or a dummy pill that looked and tasted like the aspirin but had no active ingredient (a **placebo**). On even-numbered days, they took either a capsule containing beta-carotene or a placebo. There were several response variables—the study looked for heart attacks, several kinds of cancer, and other medical outcomes. After several years, 239 of the placebo group but only 139 of the aspirin group had suffered heart attacks. This difference is large enough to give good evidence that taking aspirin does reduce heart attacks.[23] It did not appear, however, that beta-carotene had any effect on preventing cancer.

Why did researchers decide to do the Physicians' Health Study (PHS)? The interesting history that led to this experiment is detailed at the PHS Web site. You can also find out about the Physicians' Health Study II, which ended in December 2007.

PROBLEM: Explain how each of the four principles of experimental design was used in the Physicians' Health Study.

SOLUTION:

Comparison: Researchers used a design that compared both of the active treatments to a placebo.

Random assignment: Was used to determine which subjects received each of the four treatment combinations. This helped ensure that the treatment groups were roughly equivalent to begin with.

Control: The experiment used subjects of the same gender and occupation. All subjects followed the same schedule of pill taking.

Replication: There were over 5000 subjects per treatment group. This large number of subjects helped ensure that the difference in heart attacks was due to the aspirin and not to chance variation in the random assignment.

For Practice *Try Exercise* 63

The Physicians' Health Study shows how well-designed experiments can yield good evidence that differences in the treatments cause the differences we observe in the response.

Completely Randomized Designs

The diagram in Figure 4.4 presents the details of the SAT prep experiment: random assignment, the sizes of the groups and which treatment they receive, and the response variable. There are, as we will see later, statistical reasons for using treatment groups that are about equal in size. This type of design is called a **completely randomized design**.

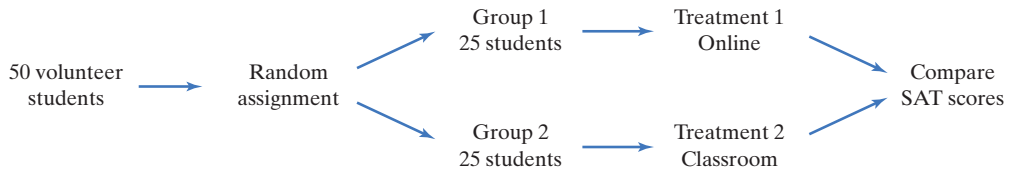


FIGURE 4.4 Outline of a completely randomized design to compare online and classroom instruction.

DEFINITION: Completely randomized design

In a **completely randomized design,** the experimental units are assigned to the treatments completely by chance.

Notice that the definition of a completely randomized design does not require that each treatment be assigned to an equal number of experimental units. It does specify that the assignment of treatments must occur completely at random.

THINK ABOUT IT

Does using chance to assign treatments in an experiment guarantee a completely randomized design? Actually, no. Let's return to the SAT prep course experiment. Another way to randomly assign the 50 students to the two treatments is by tossing a coin. Have each student come forward and toss a coin. If it's heads, then the student will take the course online. If it's tails, then the student will take the classroom course.

As long as all 50 students toss a coin, this is still a completely randomized design. Of course, the two experimental groups are unlikely to contain exactly 25 students each due to the chance variation in coin tosses.

The problem comes if we try to force the two groups to have equal sizes. Suppose we let the coin tossing continue until one of the groups has 25 students and then place the remaining students in the other group. This is no longer a completely randomized design, because the last few students aren't being assigned to one of the treatments by chance. In fact, these students will all end up in the same group, which could lead to bias if these individuals share some characteristic that would systematically affect the response variable. For example, if the students came to toss the coin last because they're lazier than the other students who volunteered, then the SAT prep class that they're in will seem less effective than it really is.

Completely randomized designs can compare any number of treatments. Here is an experiment that compares three treatments.

EXAMPLE

Conserving Energy

A completely randomized design

Many utility companies have introduced programs to encourage energy conservation among their customers. An electric company considers placing small digital displays in households to show current electricity use and what the cost would be if this use continued for a month. Will the displays reduce electricity use? One cheaper approach is to give customers a chart and information about monitoring their electricity use from their outside meter. Would this method work almost as well? The company decides to conduct an experiment to compare these two approaches (display, chart) with a group of customers who receive information about energy consumption but no help in monitoring electricity use.

PROBLEM: Describe a completely randomized design involving 60 single-family residences in the same city whose owners are willing to participate in such an experiment. Write a few sentences explaining how you would implement your design.

SOLUTION: Figure 4.5 outlines the design. We'll randomly assign 20 houses to each of three treatments: digital display, chart plus information, and information only. Our response variable is the total amount of electricity used in a year.

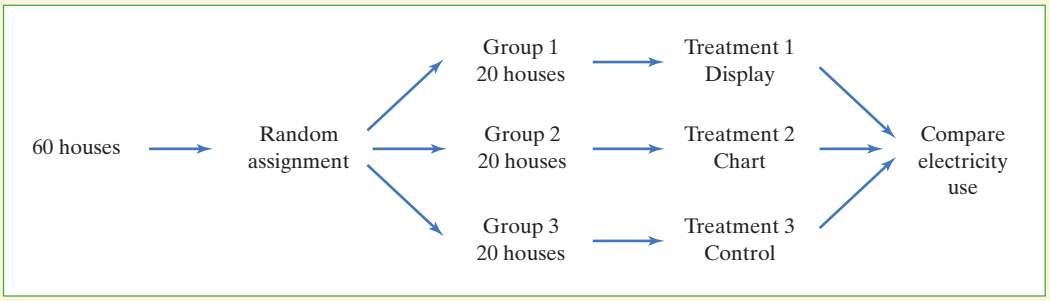


FIGURE 4.5 Outline of a completely randomized design to compare three energy-saving programs.

To implement the design, start by labeling each house with a distinct number from 1 to 60. Write the labels on 60 identical slips of paper, put them in a hat, and mix them well. Draw out 20 slips. The corresponding homes will be given digital displays showing current electricity use. Now draw out 20 more slips. Those homes will use a chart. The remaining 20 houses will be given information about energy consumption but no way to monitor their usage. At the end of the year, compare how much electricity was used by the homes in the three groups.

AP® EXAM TIP If you are asked to describe the design of an experiment on the AP® exam, you won't get full credit for providing only a diagram like Figure 4.5. You are expected to describe how the treatments are assigned to the experimental units and to clearly state what will be measured or compared. Some students prefer to start with a diagram and then add a few sentences. Others choose to skip the diagram and put their entire response in narrative form.

For Practice *Try Exercise* 65

Why did we include a **control group** of 20 houses in the energy conservation experiment? *The main purpose of a control group is to provide a baseline for comparing the effects of the other treatments.* Without such a comparison group, we wouldn't be able to tell whether homes with digital displays or charts used less electricity than homes without such aids.

THINK ABOUT IT

Was a control group really necessary? You might be thinking that the change in electricity use from last year to this year in the houses with displays and charts would tell us whether these treatments helped. Unfortunately, it's not that simple. Suppose last year's temperatures were more extreme than this year's. Then many households might show a decrease in electricity use, but we couldn't be sure whether this change was due to the weather or to the treatments. (Can you say confounding?!)

Many experiments (like the one in the previous example) include a control group that receives an inactive treatment. However, a control group can also be given an active treatment. Suppose we want to compare the effectiveness of a newly developed drug for treating a serious disease with a drug that's already known to work. In that case, the experimental units that receive the existing drug form the control group.

Some experimental designs don't include a control group. That's appropriate if researchers simply want to compare the effects of several treatments, rather than determining whether any of them works better than an inactive treatment. For instance, a state's highway department wants to see which of three brands of paint will last longest when marking lane lines on the freeway. Putting no paint on the highway is clearly not a good option!

CHECK YOUR UNDERSTANDING

Music students often don't evaluate their own performances accurately. Can small-group discussions help? The subjects were 29 students preparing for the end-of-semester performance that is an important part of their grade. The 15 students in one group each videotaped a practice performance, evaluated it themselves, and then discussed the tape with a small group of other students. The remaining 14 students watched and evaluated their tapes alone. At the end of the semester, the discussion-group students evaluated their final performance more accurately.[24]

1. Describe a completely randomized design for this experiment. Write a few sentences describing how you would implement your design.
2. What is the purpose of the control group in this experiment?

Experiments: What Can Go Wrong?

The logic of a randomized comparative experiment depends on our ability to treat all the subjects the same in every way except for the actual treatments being compared. Good experiments, therefore, require careful attention to details to ensure that all subjects really are treated identically.

If some subjects in a medical experiment take a pill each day and a control group takes no pill, the subjects are not treated identically. Many medical experiments are therefore "placebo-controlled," like the Physicians' Health Study. On odd-numbered days, all the subjects took an aspirin or a placebo. On even-numbered days, all of them took either a beta-carotene pill or a placebo.

Many patients respond favorably to any treatment, even a placebo, perhaps because they trust the doctor. The response to a dummy treatment is called the **placebo effect**. If some subjects in the Physicians' Health Study did not take any pills, the effect of aspirin or beta-carotene would be confounded with the placebo effect, the effect of simply taking pills.

EXAMPLE

Curing Baldness and Soothing Pain

Do placebos work?

Want to help balding men keep their hair? Give them a placebo. One study found that 42% of balding men maintained or increased the amount of hair on their heads when they took a placebo. In another study, researchers zapped the wrists

of 24 test subjects with a painful jolt of electricity. Then they rubbed a cream with no active medicine on subjects' wrists and told them the cream should help soothe the pain. When researchers shocked them again, 8 subjects said they experienced significantly less pain.[25]

When the ailment is vague and psychological, like depression, some experts think that the placebo effect accounts for about three-quarters of the effect of the most widely used drugs.[26] Others disagree. In any case, "placebos work" is a good place to start when you think about planning medical experiments.

The strength of the placebo effect is a strong argument for randomized comparative experiments. In the baldness study, 42% of the placebo group kept or increased their hair, but 86% of the men getting a new drug to fight baldness did so. The drug beats the placebo, so it has something besides the placebo effect going for it. Of course, the placebo effect is still part of the reason this and other treatments work.

Because the placebo effect is so strong, it would be foolish to tell subjects in a medical experiment whether they are receiving a new drug or a placebo. Knowing that they are getting "just a placebo" might weaken the placebo effect and bias the experiment in favor of the other treatments.

It is also foolish to tell doctors and other medical personnel what treatment each subject received. If they know that a subject is getting "just a placebo," they may expect less than if they know the subject is receiving a promising experimental drug. Doctors' expectations change how they interact with patients and even the way they diagnose a patient's condition. Whenever possible, experiments with human subjects should be **double-blind**.

Reports in medical journals regularly begin with words like these, from a study of a flu vaccine given as a nose spray: "This study was a randomized, double-blind, placebo-controlled trial. Participants were enrolled from 13 sites across the continental United States between mid-September and mid-November."[27] Doctors are supposed to know what this means. Now you know, too.

DEFINITION: Double-blind

In a **double-blind** experiment, neither the subjects nor those who interact with them and measure the response variable know which treatment a subject received.

The idea of a double-blind design is simple. Until the experiment ends and the results are in, only the study's statistician knows for sure which treatment a subject is receiving. However, some experiments cannot be carried out in a double-blind manner. If researchers are comparing the effects of exercise and dieting on weight loss, then subjects will know which treatment they are receiving. Such an experiment can still be **single-blind** if the individuals who are interacting with the subjects and measuring the response variable don't know who is dieting and who is exercising. In other single-blind experiments, the subjects are unaware of which treatment they are receiving, but the people interacting with them and measuring the response variable do know.

CHECK YOUR UNDERSTANDING

In an interesting experiment, researchers examined the effect of ultrasound on birth weight. Pregnant women participating in the study were randomly assigned to one of two groups. The first group of women received an ultrasound; the second group did not. When the subjects' babies were born, their birth weights were recorded. The women who received the ultrasounds had heavier babies.[28]

1. Did the experimental design take the placebo effect into account? Why is this important?
2. Was the experiment double-blind? Why is this important?
3. Based on your answers to Questions 1 and 2, describe an improved design for this experiment.

Inference for Experiments

In an experiment, researchers usually hope to see a difference in the responses so large that it is unlikely to happen just because of chance variation. We can use the laws of probability, which describe chance behavior, to learn whether the treatment effects are larger than we would expect to see if only chance were operating. If they are, we call them **statistically significant**.

DEFINITION: Statistically significant

An observed effect so large that it would rarely occur by chance is called **statistically significant.**

If we observe statistically significant differences among the groups in a randomized comparative experiment, we have good evidence that the treatments caused these differences. You will often see the phrase "statistically significant" in published research reports in many fields. The great advantage of randomized comparative experiments is that they can produce data that give good evidence for a cause-and-effect relationship between the explanatory and response variables. We know that in general a strong association does not imply causation. A statistically significant association in data from a well-designed experiment *does* imply causation.

ACTIVITY | Distracted driving

MATERIALS: Set of 48 index cards or standard deck of playing cards for each pair of students

Is talking on a cell phone while driving more distracting than talking to a passenger? David Strayer and his colleagues at the University of Utah designed an experiment to help answer this question. They used 48 undergraduate students as subjects. The researchers randomly assigned half of the subjects to drive in a simulator while talking on a cell phone, and the other half to drive in the simulator

while talking to a passenger. One response variable was whether or not the driver stopped at a rest area that was specified by researchers before the simulation started. The table below shows the results:[29]

	Distraction	
Stopped at rest area?	**Cell phone**	**Passenger**
Yes	12	21
No	12	3

Are these results statistically significant? To find out, let's see what would happen just by chance if we randomly reassign the 48 people in this experiment to the two groups many times, *assuming the treatment received doesn't affect whether a driver stops at the rest area*.

1. We need 48 cards to represent the drivers in this study. In the original experiment, 33 drivers stopped at the rest area and 15 didn't. Because we're assuming that the treatment received won't change whether each driver stops at the rest area, we use 33 cards to represent drivers who stop and 15 cards to represent those who don't.
   - *Using index cards:* Write "Yes" on 33 cards and "No" on 15 cards.
   - *Using playing cards:* Remove jokers and other specialty cards from the deck, as well as the ace of spades and any three of the 2s. All cards with denominations 2 through 10 represent drivers who stop. All jacks, queens, kings, and aces represent drivers who don't stop.

Why did we only ask you to count the number of drivers who didn't stop at the rest area in the cell-phone group? Suppose you get 10 in one trial of the simulation. That means the other 24 − 10 drivers in the cell-phone group did stop at the rest area. Also, there must be 15 − 10 = 5 drivers in the passenger group who failed to stop, leaving 24 − 5 = 19 drivers in the group who did stop. Recording the one number tells you all the others.

2. Shuffle and deal two piles of 24 cards each—the first pile represents the cell phone group and the second pile represents the passenger group. The shuffling reflects our assumption that the outcome for each subject is not affected by the treatment. Record the number of drivers who failed to stop at the rest area in the cell-phone group.

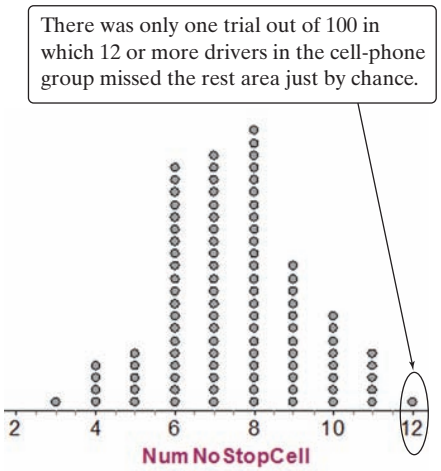
3. Your teacher will draw and label axes for a class dotplot. Plot the result you got in Step 2 on the graph.

4. Repeat Steps 2 and 3 if needed to get a total of at least 40 repetitions of the simulation for your class.

5. In the original experiment, 12 of the 24 drivers using cell phones didn't stop at the rest area. Based on the class's simulation results, how surprising would it be to get a result this large or larger simply due to the chance involved in the random assignment? Is the result statistically significant?

6. What conclusion would you draw about whether talking on a cell phone is more distracting than talking to a passenger?

Here is an example of what the class dotplot in the Activity might look like after 100 trials. In the 100 trials, only once did 12 or more people fail to stop when using a cell phone. Because a result of 12 or more is unlikely to happen by chance alone, the results of this study should be considered statistically significant.



THINK ABOUT IT

Can an "unlucky" random assignment lead to confounding?

Let's return to the distracted-driver Activity. Some people are more forgetful than others. Suppose that the random assignment happens to put most of the forgetful subjects in one group. If more drivers in that group fail to stop at the rest area, we don't know if it's because of the treatment they received (cell phone or passenger) or their forgetfulness. Is this confounding?

You might be surprised that the answer is "No!" Although people's memory is a variable that might affect whether or not they stop at the rest area (the response variable), the design of the experiment takes care of this by randomly assigning subjects to the two treatment groups. The "unlucky" random assignments are taken into account in determining statistical significance. In an experiment, confounding occurs when the design doesn't account for existing differences in the experimental units that might systematically affect their response to the treatments.

Blocking

Completely randomized designs are the simplest statistical designs for experiments. They illustrate clearly the principles of comparison, random assignment, control, and replication. But just as with sampling, there are times when the simplest method doesn't yield the most precise results. When a population consists of groups of individuals that are "similar within but different between," a stratified random sample gives a better estimate than a simple random sample. This same logic applies in experiments.

EXAMPLE

A Smarter Design?

The idea of blocking

Suppose that a mobile phone company is considering two different keyboard designs (A and B) for its new smart phone. The company decides to perform an experiment to compare the two keyboards using a group of 10 volunteers. The response variable is typing speed, measured in words per minute.

How should the company deal with the fact that four of the volunteers already use a smart phone, whereas the remaining six volunteers do not? They could use a completely randomized design and hope that the random assignment distributes the smart-phone users and non-smart-phone users about evenly between the group using keyboard A and the group using keyboard B. Even so, there might be a lot of variability in typing speed in both groups because some members of each group are much more familiar with smart phones than others. This additional variability might make it difficult to detect a difference in the effectiveness of the two keyboards. What should the researchers do?

Because the company knows that experience with smart phones will affect typing speed, they could start by separating the volunteers into two groups—one with experienced smart-phone users and one with inexperienced smart-phone users. Each of these groups of similar subjects is known as a **block**. Within each block, the company could then randomly assign half of the subjects to use keyboard A and the other half to use keyboard B. To control other variables, each subject should be given the same passage to type while in a quiet room with no distractions. This **randomized block design** helps account for the variation in typing speed that is due to experience with smart phones.

Figure 4.6 outlines the randomized block design for the smart-phone experiment. The subjects are first separated into blocks based on their experience with smart phones. Then the two treatments are randomly assigned within each block.

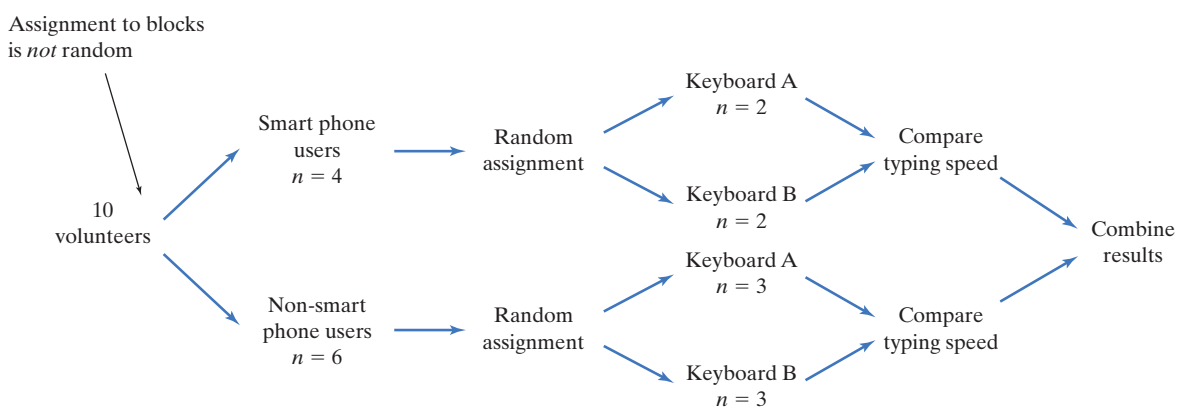


Figure 4.6 Outline of a randomized block design for the smart-phone experiment. The blocks consist of volunteers who have used smart phones and volunteers who have not used smart phones. The treatments are keyboard A and keyboard B.

DEFINITION: Block and randomized block design

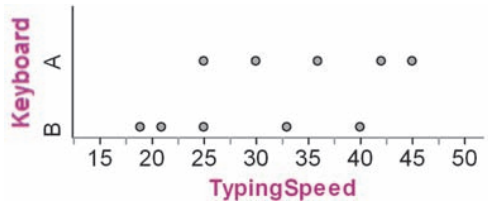
A **block** is a group of experimental units that are known before the experiment to be similar in some way that is expected to affect the response to the treatments.

In a **randomized block design**, the random assignment of experimental units to treatments is carried out separately within each block.

Using a randomized block design allows us to account for the variation in the response that is due to the blocking variable. This makes it easier to determine if one treatment is really more effective than the other.

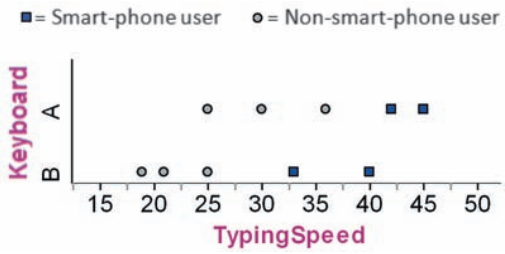
To see how blocking helps, let's look at the results of an experiment using 10 volunteers, 4 who already use a smart phone and 6 who do not. In the block of 4 smart-phone users, 2 will be randomly assigned to use keyboard A and the other 2 will be assigned to use keyboard B. Likewise, in the block of 6 non-smart-phone users, 3 will be randomly assigned to use keyboard A and the other 3 will be assigned to use keyboard B. Each of the 10 volunteers will type the same passage and the typing speed will be recorded.

Here are the results:



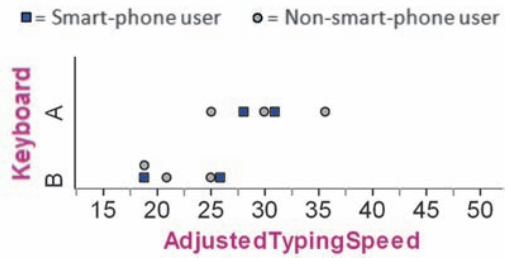
There is some evidence that keyboard A results in higher typing speeds, but the evidence isn't that convincing. Enough overlap occurs in the two distributions that the differences might simply be due to the chance variation in the random assignment.

If we compare the results for the two keyboards within each block, however, a different story emerges. Among the 4 smart-phone users (indicated by the blue squares), keyboard A was the clear winner. Likewise, among the 6 non-smart-phone users (indicated by the gray dots), keyboard A was also the clear winner.



The overlap in the first set of dotplots was due almost entirely to the variation in smart-phone experience—smart-phone users were generally faster than non-smart-phone users, regardless of which keyboard they used. In fact, the average typing speed for the smart-phone users was 40, while the average typing speed for non-smart-phone users was only 26, a difference of 14 words per minute. To account for the variation created by the difference in smart-phone experience, let's subtract 14 from each of the typing speeds in the block of smart-phone users to "even the playing field."

Here are the results:



AP® EXAM TIP Don't mix the language of experiments and the language of sample surveys or other observational studies. You will lose credit for saying things like "use a randomized block design to select the sample for this survey" or "this experiment suffers from nonresponse since some subjects dropped out during the study."

Because we accounted for the variation due to the difference in smart-phone experience, the variation in each of the distributions has been reduced. There is now almost no overlap between the two distributions, meaning that the evidence in favor of keyboard A is much more convincing. *When blocks are formed wisely, it is easier to find convincing evidence that one treatment is more effective than another.*

The idea of blocking is an important additional principle of experimental design. A wise experimenter will form blocks based on the most important unavoidable sources of variation (other variables) among the experimental units. Randomization will then average out the effects of the remaining other variables and allow an unbiased comparison of the treatments. The moral of the story is: *control what you can, block on what you can't control, and randomize to create comparable groups.*

EXAMPLE Men, Women, and Advertising

Blocking in an experiment

Women and men respond differently to advertising. Researchers would like to design an experiment to compare the effectiveness of three advertisements for the same product.

PROBLEM:

(a) Explain why a randomized block design might be preferable to a completely randomized design for this experiment.

(b) Outline a randomized block design using 300 volunteers (180 men and 120 women) as subjects. Describe how you would carry out the random assignment required by your design.

SOLUTION:

(a) A completely randomized design considers all subjects, both men and women, as a single pool. The random assignment would send subjects to three treatment groups without regard to their gender. This ignores the differences between men and women, which would probably result in a great deal of variability in responses to the advertising in all three groups. For example, if an ad appealed much more to men, you would get a wide range of reactions to that ad from the two genders. That would make it harder to determine whether one ad was more effective.

A randomized block design would consider women and men separately. In this case, the random assignment would occur separately in each block. Blocking will account for the variability in responses to advertising due to gender. This will allow researchers to look separately at the reactions of men and women, as well as to more effectively assess the overall response to the ads.

(b) Figure 4.7 outlines the randomized block design. We randomly assign the 120 women into three groups of 40, one for each of the advertising treatments. Write the women's names on 120 identical slips of paper, place the slips in a hat, and mix them well. Pull out 40 slips to determine which women will view Ad 1. Pull out another 40 slips to determine which women will view Ad 2. The remaining 40 women will view Ad 3. Randomly assign the 180 men into three groups of 60 using a similar process. After each subject has viewed the assigned ad, compare reactions to the three ads within the gender blocks. To compare the overall effectiveness of the three ads, combine the results from the two blocks after accounting for the difference in response for the men and women.

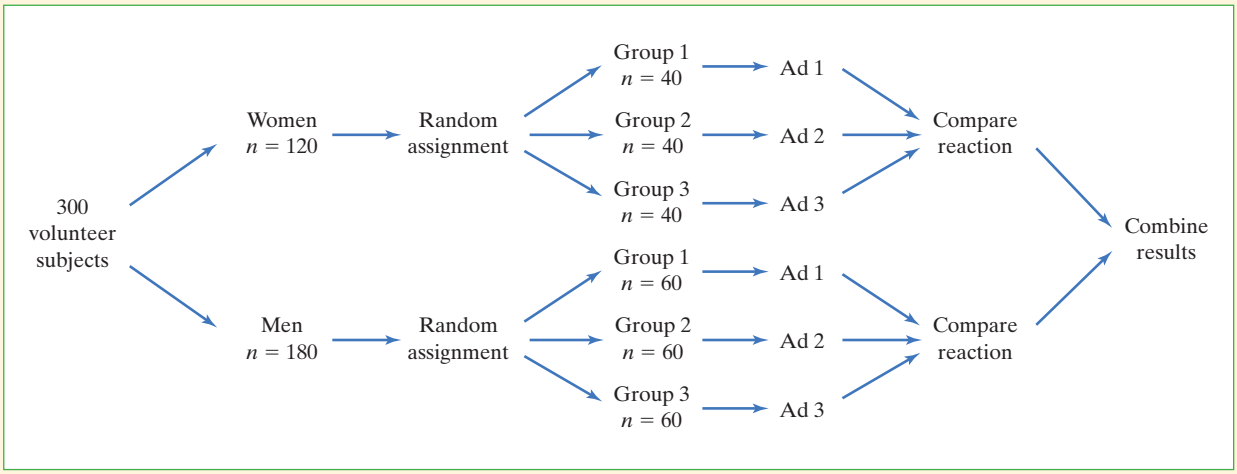


FIGURE 4.7 Randomized block design for comparing responses to three advertisements. The blocks consist of male and female subjects.

For Practice *Try Exercise* 77

Matched Pairs Design: A common type of randomized block design for comparing two treatments is a **matched pairs design**. The idea is to create blocks by matching pairs of similar experimental units. Then we can use chance to decide which member of a pair gets the first treatment. The other subject in that pair receives the other treatment. That is, the random assignment of subjects to treatments is done within each matched pair. Just as with other forms of blocking, matching helps account for the variation among the experimental units.

Sometimes each "pair" in a matched pairs design consists of just one experimental unit that gets both treatments one after the other. In that case, each experimental unit serves as its own control. The *order* of the treatments can influence the response, so we randomize the order for each experimental unit.

ACTIVITY | Get your heart beating

MATERIALS:
Clock or stopwatch

Are standing pulse rates generally higher than sitting pulse rates? In this Activity, you will perform two experiments to try to answer this question.

1. *Completely randomized design* For the first experiment, your teacher will randomly assign half of the students in your class to stand and the other half to sit. Once the two treatment groups have been formed, students should stand or sit as required. Then they should measure their pulses for one minute. Have the subjects in each group record their data on the board.

2. *Matched pairs design* In a matched pairs design, each student should receive both treatments in a random order. Because you already sat or stood in Step 1, you just need to do the opposite now. As before, everyone should measure their

pulses for one minute after completing the treatment (that is, once they are standing or sitting). Have all the subjects record their data (both measurements) in a chart on the board.

3. Analyze the data for the completely randomized design. Make parallel dotplots and calculate the mean pulse rate for each group. Is there convincing evidence that standing pulse rates are higher? Explain.

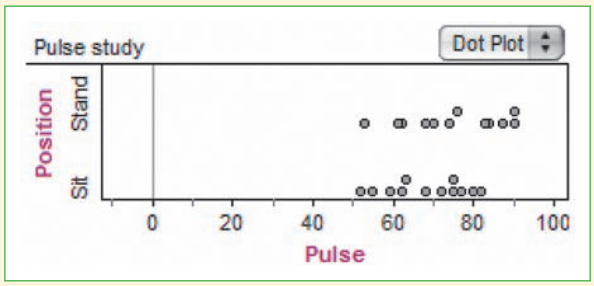
4. Analyze the data for the matched pairs design. Because the data are paired by student, your first step should be to calculate the difference in pulse rate (standing – sitting) for each subject. Make a dotplot of these differences and calculate their mean. Is there convincing evidence that standing pulse rates are higher? Explain.

5. What advantage does the matched pairs design have over the completely randomized design?

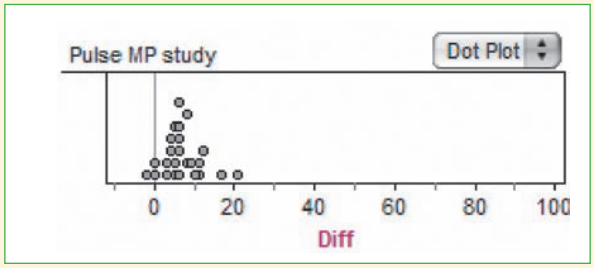
An AP® Statistics class with 24 students performed the "Get Your Heart Beating" Activity. We'll analyze the results of their experiment in the following example.

EXAMPLE Standing and Sitting Pulse Rate

Design determines analysis



A Fathom dotplot of the pulse rates for their completely randomized design is shown. The mean pulse rate for the standing group is 74.83; the mean for the sitting group is 68.33. So the average pulse rate is 6.5 beats per minute higher in the standing group. However, the variability in pulse rates for the two groups creates a lot of overlap in the dotplots. These data don't provide convincing evidence that standing pulse rates tend to be higher.



What about the class's matched pairs experiment? The Fathom dotplot shows their data on the difference in pulse rates (standing – sitting). For these 24 students, the mean difference was 6.8 beats per minute. In addition, 21 of the 24 students recorded a positive difference (meaning the standing pulse rate was higher). These data provide convincing evidence that people's standing pulse rates tend to be higher than their sitting pulse rates.

Let's take one more look at the two Fathom dotplots in the example. Notice that we used the same scale for both graphs. This is to help you visually compare the amount of variability in the response variable for each of the two experimental designs. Blocking by subject in the matched pairs design greatly reduced the

variability in the response variable. That made it easier to detect the fact that standing causes an increase in pulse rate. With the large amount of variability in the completely randomized design, we were unable to draw such a conclusion.

Another important lesson to take away from the example is this: *the design of the study determines the appropriate method of analysis*. For the completely randomized design, it makes sense to compare pulse rates for the two groups with parallel dotplots and means. In the matched pairs design, each student is a block. We compare the effects of the treatments within each block by examining the differences in standing and sitting pulse rates for each student. Then we combine the results from each block (student) and examine the distribution of differences.

The following Data Exploration asks you to apply what you have learned about analyzing data from an experiment.

DATA EXPLORATION

Nitrogen in tires–a lot of hot air?

Most automobile tires are inflated with compressed air, which consists of about 78% nitrogen. Aircraft tires are filled with pure nitrogen, which is safer than air in case of fire. Could filling automobile tires with nitrogen improve safety, performance, or both?

Consumers Union designed a study to test whether nitrogen-filled tires would maintain pressure better than air-filled tires. They obtained two tires from each of several brands and then filled one tire in each pair with air and one with nitrogen. All tires were inflated to a pressure of 30 pounds per square inch and then placed outside for a year. At the end of the year, Consumers Union measured the pressure in each tire. The amount of pressure lost (in pounds per square inch) during the year for the air-filled and nitrogen-filled tires of each brand is shown in the table below.[30]

Brand	Air	Nitrogen	Brand	Air	Nitrogen
BF Goodrich Traction T/A HR	7.6	7.2	Pirelli P6 Four Seasons	4.4	4.2
Bridgestone HP50 (Sears)	3.8	2.5	Sumitomo HTR H4	1.4	2.1
Bridgestone Potenza G009	3.7	1.6	Yokohama Avid H4S	4.3	3.0
Bridgestone Potenza RE950	4.7	1.5	BF Goodrich Traction T/A V	5.5	3.4
Bridgestone Potenza EL400	2.1	1.0	Bridgestone Potenza RE950	4.1	2.8
Continental Premier Contact H	4.9	3.1	Continental ContiExtreme Contact	5.0	3.4
Cooper Lifeliner Touring SLE	5.2	3.5	Continental ContiProContact	4.8	3.3
Dayton Daytona HR	3.4	3.2	Cooper Lifeliner Touring SLE	3.2	2.5
Falken Ziex ZE-512	4.1	3.3	General Exclaim UHP	6.8	2.7
Fuzion Hrl	2.7	2.2	Hankook Ventus V4 H105	3.1	1.4
General Exclaim	3.1	3.4	Michelin Energy MXV4 Plus	2.5	1.5
Goodyear Assurance Tripletred	3.8	3.2	Michelin Pilot Exalto A/S	6.6	2.2
Hankook Optimo H418	3.0	0.9	Michelin Pilot HX MXM4	2.2	2.0
Kumho Solus KH16	6.2	3.4	Pirelli P6 Four Seasons	2.5	2.7
Michelin Energy MXV4 Plus	2.0	1.8	Sumitomo HTR^{+}	4.4	3.7
Michelin Pilot XGT H4	1.1	0.7			

Does filling tires with nitrogen instead of compressed air reduce pressure loss? Give appropriate graphical and numerical evidence to support your answer.

Section 4.2 Summary

- We can produce data intended to answer specific questions by **observational studies** or **experiments**. An observational study gathers data on individuals as they are. Experiments actively do something to people, animals, or objects in order to measure their response.
- Statistical studies often try to show that changing one variable (the **explanatory variable**) causes changes in another variable (the **response variable**). Variables are **confounded** when their effects on a response variable can't be distinguished from each other. Observational studies and uncontrolled experiments often fail to show that changes in an explanatory variable cause changes in a response variable because the explanatory variable is confounded with other variables.
- In an experiment, we impose one or more **treatments** on a group of **experimental units** (sometimes called **subjects** if they are human). Each treatment is a combination of values of the explanatory variables (also called **factors**).
- The basic principles of experimental design are as follows:
  1. *Comparison:* Use a design that compares two or more treatments.
  2. *Random assignment:* Use chance to assign experimental units to treatments. This helps create roughly equivalent groups before treatments are imposed.
  3. *Control:* Keep as many other variables as possible the same for all groups. Control helps avoid confounding and reduces the variation in responses, making it easier to decide whether a treatment is effective.
  4. *Replication:* Impose each treatment on enough experimental units so that the effects of the treatments can be distinguished from chance differences between the groups.
- In a **completely randomized design**, all of the experimental units are assigned to the treatments completely by chance.
- Some experiments give a **placebo** (fake treatment) to a control group. That helps prevent confounding due to the **placebo effect**, in which some patients get better because they expect the treatment to work.
- Many behavioral and medical experiments are **double-blind**. That is, neither the subjects nor those interacting with them and measuring their responses know who is receiving which treatment. If one party knows and the other doesn't, then the experiment is **single-blind**.
- When an observed difference in responses between the groups in an experiment is too large to be explained by chance variation in the random assignment, we say that the result is **statistically significant**.
- A **randomized block design** forms groups (**blocks**) of experimental units that are similar with respect to a variable that is expected to affect the response. Treatments are assigned at random within each block. Responses are then compared within each block and combined with the responses of other blocks after accounting for the differences between the blocks. When blocks are chosen wisely, it is easier to determine if one treatment is more effective than another.

- A **matched pairs design** is a common form of blocking for comparing just two treatments. In some matched pairs designs, each subject receives both treatments in a random order. In others, two very similar subjects are paired, and the two treatments are randomly assigned within each pair.

Section 4.2 Exercises

45. **Learning biology with computers** An educator wants to compare the effectiveness of computer software for teaching biology with that of a textbook presentation. She gives a biology pretest to each of a group of high school juniors, then randomly divides them into two groups. One group uses the computer, and the other studies the text. At the end of the year, she tests all the students again and compares the increase in biology test scores in the two groups. Is this an observational study or an experiment? Justify your answer.

46. **Cell phones and brain cancer** One study of cell phones and the risk of brain cancer looked at a group of 469 people who have brain cancer. The investigators matched each cancer patient with a person of the same age, gender, and race who did not have brain cancer, then asked about the use of cell phones. Result: "Our data suggest that the use of handheld cellular phones is not associated with risk of brain cancer."[31] Is this an observational study or an experiment? Justify your answer.

47. **Chocolate and happy babies** A University of Helsinki (Finland) study wanted to determine if chocolate consumption during pregnancy had an effect on infant temperament at age 6 months. Researchers began by asking 305 healthy pregnant women to report their chocolate consumption. Six months after birth, the researchers asked mothers to rate their infants' temperament, including smiling, laughter, and fear. The babies born to women who had been eating chocolate daily during pregnancy were found to be more active and "positively reactive"—a measure that the investigators said encompasses traits like smiling and laughter.[32]

(a) Was this an observational study or an experiment? Justify your answer.

(b) What are the explanatory and response variables?

(c) Does this study show that eating chocolate regularly during pregnancy helps produce infants with good temperament? Explain.

48. **Child care and aggression** A study of child care enrolled 1364 infants and followed them through their sixth year in school. Later, the researchers published an article in which they stated that "the more time children spent in child care from birth to age four-and-a-half, the more adults tended to rate them, both at age four-and-a-half and at kindergarten, as less likely to get along with others, as more assertive, as disobedient, and as aggressive."[33]

(a) Is this an observational study or an experiment? Justify your answer.

(b) What are the explanatory and response variables?

(c) Does this study show that child care causes children to be more aggressive? Explain.

49. **Effects of class size** Do smaller classes in elementary school really benefit students in areas such as scores on standardized tests, staying in school, and going on to college? We might do an observational study that compares students who happened to be in smaller and larger classes in their early school years. Identify a variable that may lead to confounding with the effects of small classes. Explain how confounding might occur.

50. **Effects of binge drinking** A common definition of "binge drinking" is 5 or more drinks at one sitting for men and 4 or more for women. An observational study finds that students who binge drink have lower average GPA than those who don't. Identify a variable that may be confounded with the effects of binge drinking. Explain how confounding might occur.

For the experiments described in Exercises 51 to 56, identify the experimental units, the explanatory and response variables, and the treatments.

pg 237

51. **Growing in the shade** Ability to grow in shade may help pines found in the dry forests of Arizona to resist drought. How well do these pines grow in shade? Investigators planted pine seedlings in a greenhouse in either full light, light reduced to 25% of normal by shade cloth, or light reduced to 5% of normal. At the end of the study, they dried the young trees and weighed them.

52. **Internet telephone calls** You can use Voice over Internet Protocol (VoIP) to make long-distance calls over the Internet. One of the most popular VoIP services is Skype. How will the appearance of ads during calls affect the use of this service? Researchers design an experiment to find out. They recruit 300 people who

have not used Skype before to participate. Some people get the current version of Skype with no ads. Others see ads whenever they make calls. The researchers are interested in frequency and length of phone calls.

53. **Improving response rate** How can we reduce the rate of refusals in telephone surveys? Most people who answer at all listen to the interviewer's introductory remarks and then decide whether to continue. One study made telephone calls to randomly selected households to ask opinions about the next election. In some calls, the interviewer gave her name; in others, she identified the university she was representing; and in still others, she identified both herself and the university. For each type of call, the interviewer either did or did not offer to send a copy of the final survey results to the person interviewed. Do these differences in the introduction affect whether the interview is completed?

54. **Eat well and exercise** Most American adolescents don't eat well and don't exercise enough. Can middle schools increase physical activity among their students? Can they persuade students to eat better? Investigators designed a "physical activity intervention" to increase activity in physical education classes and during leisure periods throughout the school day. They also designed a "nutrition intervention" that improved school lunches and offered ideas for healthy home-packed lunches. Each participating school was randomly assigned to one of the interventions, both interventions, or no intervention. The investigators observed physical activity and lunchtime consumption of fat.

55. **Fabric science** A maker of fabric for clothing is setting up a new line to "finish" the raw fabric. The line will use either metal rollers or natural-bristle rollers to raise the surface of the fabric; a dyeing-cycle time of either 30 or 40 minutes; and a temperature of either 150° or 175°C. An experiment will compare all combinations of these choices. Three specimens of fabric will be subjected to each treatment and scored for quality.

pg 239

56. **Exercise and heart rate** A student project measured the increase in the heart rates of fellow students when they stepped up and down for 3 minutes to the beat of a metronome. The step was either 5.75 or 11.5 inches high and the metronome beat was 14, 21, or 28 steps per minute. Thirty students took part in the experiment. Five of them stepped at each combination of height and speed.

57. **Cocoa and blood flow** A study conducted by Norman Hollenberg, professor of medicine at Brigham and Women's Hospital and Harvard Medical School, involved 27 healthy people aged 18 to 72. Each subject consumed a cocoa beverage containing 900 milligrams of flavonols (a class of flavonoids) daily for 5 days. Using a finger cuff, blood flow was measured on the first and fifth days of the study. After 5 days, researchers measured what they called "significant improvement" in blood flow and the function of the cells that line the blood vessels.[34] What flaw in the design of this experiment makes it impossible to say whether the cocoa really caused the improved blood flow? Explain.

58. **Reducing unemployment** Will cash bonuses speed the return to work of unemployed people? A state department of labor notes that last year 68% of people who filed claims for unemployment insurance found a new job within 15 weeks. As an experiment, this year the state offers $500 to people filing unemployment claims if they find a job within 15 weeks. The percent who do so increases to 77%. What flaw in the design of this experiment makes it impossible to say whether the bonus really caused the increase? Explain.

59. **Layoffs and "survivor guilt"** Workers who survive a layoff of other employees at their location may suffer from "survivor guilt." A study of survivor guilt and its effects used as subjects 120 students who were offered an opportunity to earn extra course credit by doing proofreading. Each subject worked in the same cubicle as another student, who was an accomplice of the experimenters. At a break midway through the work, one of three things happened:

pg 241

Treatment 1: The accomplice was told to leave; it was explained that this was because she performed poorly.

Treatment 2: It was explained that unforeseen circumstances meant there was only enough work for one person. By "chance," the accomplice was chosen to be laid off.

Treatment 3: Both students continued to work after the break.

The subjects' work performance after the break was compared with performance before the break.[35] Describe how you would randomly assign the subjects to the treatments

(a) using slips of paper.
(b) using technology.
(c) using Table D.

60. **Effects of TV advertising** Figure 4.2 (page 239) displays the 6 treatments for a two-factor experiment on TV advertising. Suppose we have 150 students who are willing to serve as subjects. Describe how you would randomly assign the subjects to the treatments

(a) using slips of paper.
(b) using technology.
(c) using Table D.

61. **Stronger players** A football coach hears that a new exercise program will increase upper-body strength better than lifting weights. He is eager to test this new program in the off-season with the players on his high school team. The coach decides to let his players choose which of the two treatments they will undergo for 3 weeks—exercise or weight lifting. He will use the

number of push-ups a player can do at the end of the experiment as the response variable. Which principle of experimental design does the coach's plan violate? Explain how this violation could lead to confounding.

62. **Killing weeds** A biologist would like to determine which of two brands of weed killer, X or Y, is less likely to harm the plants in a garden at the university. Before spraying near the plants, the biologist decides to conduct an experiment using 24 individual plants. Which of the following two plans for randomly assigning the treatments should the biologist use? Why?

Plan A: Choose the 12 healthiest-looking plants. Then flip a coin. If it lands heads, apply Brand X weed killer to these plants and Brand Y weed killer to the remaining 12 plants. If it lands tails, do the opposite.

Plan B: Choose 12 of the 24 plants at random. Apply Brand X weed killer to those 12 plants and Brand Y weed killer to the remaining 12 plants.

63. **Do diets work?** Dr. Linda Stern and her colleagues recruited 132 obese adults at the Philadelphia Veterans Affairs Medical Center in Pennsylvania. Half the participants were randomly assigned to a low-carbohydrate diet and the other half to a low-fat diet. Researchers measured each participant's change in weight and cholesterol level after six months and again after one year. Explain how each of the four principles of experimental design was used in this study.

pg 243

64. **The effects of day care** Does day care help low-income children stay in school and hold good jobs later in life? The Carolina Abecedarian Project (the name suggests the ABCs) has followed a group of 111 children since 1972. Back then, these individuals were all healthy but low-income black infants in Chapel Hill, North Carolina. All the infants received nutritional supplements and help from social workers. Half were also assigned at random to an intensive preschool program.[36] Explain how each of the four principles of experimental design was used in this study.

65. **Headache relief** Doctors identify "chronic tension-type headaches" as headaches that occur almost daily for at least six months. Can antidepressant medications or stress management training reduce the number and severity of these headaches? Are both together more effective than either alone? Researchers want to compare four treatments: antidepressant alone, placebo alone, antidepressant plus stress management, and placebo plus stress management. Describe a completely randomized design involving 36 headache sufferers who are willing to participate in this experiment. Write a few sentences describing how you would implement your design.

pg 245

66. **More rain for California?** The changing climate will probably bring more rain to California, but we don't know whether the additional rain will come during the winter wet season or extend into the long dry season in spring and summer. Kenwyn Suttle of the University of California at Berkeley and his coworkers wanted to compare the effects of three treatments: added water equal to 20% of annual rainfall either during January to March (winter) or during April to June (spring), and no added water (control). Eighteen plots of open grassland, each with area 70 square meters, were available for this study. One response variable was total plant biomass, in grams per square meter, produced in a plot over a year.[37]

Describe a completely randomized design for this experiment. Write a few sentences describing how you would implement your design.

67. **Treating prostate disease** A large study used records from Canada's national health care system to compare the effectiveness of two ways to treat prostate disease. The two treatments are traditional surgery and a new method that does not require surgery. The records described many patients whose doctors had chosen each method. The study found that patients treated by the new method were significantly more likely to die within 8 years.[38]

(a) Further study of the data showed that this conclusion was wrong. The extra deaths among patients who got the new method could be explained by other variables. What other variables might be confounded with a doctor's choice of surgical or nonsurgical treatment?

(b) You have 300 prostate patients who are willing to serve as subjects in an experiment to compare the two methods. Describe a completely randomized design for this experiment. Write a few sentences describing how you would implement your design.

68. **Getting teachers to come to school** Elementary schools in rural India are usually small, with a single teacher. The teachers often fail to show up for work. Here is an idea for improving attendance: give the teacher a digital camera with a tamperproof time and date stamp and ask a student to take a photo of the teacher and class at the beginning and end of the day. Offer the teacher better pay for good attendance, verified by the photos. Will this work? Researchers obtained permission to use 120 rural schools in Rajasthan for an experiment to find out.[39]

(a) Explain why it would not be a good idea to offer better pay for good attendance to the teachers in all 120 schools and then to compare this year's attendance with last year's.

(b) Describe a completely randomized design for an experiment involving these 120 schools. Write a few sentences describing how you would implement your design.

69. **Do placebos really work?** Researchers in Japan conducted an experiment on 13 individuals who were extremely allergic to poison ivy. On one arm,

each subject was rubbed with a poison ivy leaf and told the leaf was harmless. On the other arm, each subject was rubbed with a harmless leaf and told it was poison ivy. All the subjects developed a rash on the arm where the harmless leaf was rubbed. Of the 13 subjects, 11 did not have any reaction to the real poison ivy leaf.[40] Explain how the results of this study support the idea of a placebo effect.

70. **Pain relief study** Fizz Laboratories, a pharmaceutical company, has developed a new drug for relieving chronic pain. Sixty patients suffering from arthritis and needing pain relief are available. Each patient will be treated and asked an hour later, "About what percent of pain relief did you experience?"

(a) Why should Fizz not simply give the new drug to 30 patients and no treatment to the other 30 patients, and then record the patients' responses?

(b) Should the patients be told whether they are getting the new drug or a placebo? How would this knowledge probably affect their reactions?

71. **Meditation for anxiety** An experiment that claimed to show that meditation lowers anxiety proceeded as follows. The experimenter interviewed the subjects and rated their level of anxiety. Then the subjects were randomly assigned to two groups. The experimenter taught one group how to meditate and they meditated daily for a month. The other group was simply told to relax more. At the end of the month, the experimenter interviewed all the subjects again and rated their anxiety level. The meditation group now had less anxiety. Psychologists said that the results were suspect because the ratings were not blind. Explain what this means and how lack of blindness could bias the reported results.

72. **Testosterone for older men** As men age, their testosterone levels gradually decrease. This may cause a reduction in lean body mass, an increase in fat, and other undesirable changes. Do testosterone supplements reverse some of these effects? A study in the Netherlands assigned 237 men aged 60 to 80 with low or low-normal testosterone levels to either a testosterone supplement or a placebo. The report in the *Journal of the American Medical Association* described the study as a "double-blind, randomized, placebo-controlled trial."[41] Explain each of these terms to someone who knows nothing about statistics.

73. **Do diets work?** Refer to Exercise 63. Subjects in the low-carb diet group lost significantly more weight than subjects in the low-fat diet group during the first six months. At the end of a year, however, the average weight loss for subjects in the two groups was not significantly different.[42]

(a) Why did researchers randomly assign the subjects to the diet treatments?

(b) Explain to someone who knows little statistics what "lost significantly more weight" means.

(c) The subjects in the low-carb diet group lost an average of 5.1 kg in a year. The subjects in the low-fat diet group lost an average of 3.1 kg. Explain how this information could be consistent with the fact that weight loss in the two groups was not significantly different.

74. **Acupuncture and pregnancy** A study sought to determine whether the ancient Chinese art of acupuncture could help infertile women become pregnant.[43] One hundred sixty healthy women undergoing assisted reproductive therapy were recruited for the study. Half of the subjects were randomly assigned to receive acupuncture treatment 25 minutes before embryo transfer and again 25 minutes after the transfer. The remaining 80 subjects were instructed to lie still for 25 minutes after the embryo transfer. *Results:* In the acupuncture group, 34 women became pregnant. In the control group, 21 women became pregnant.

(a) Why did researchers randomly assign the subjects to the two treatments?

(b) The difference in the percent of women who became pregnant in the two groups is statistically significant. Explain what this means to someone who knows little statistics.

(c) Explain why the design of the study prevents us from concluding that acupuncture caused the difference in pregnancy rates.

75. **Doctors and nurses** Nurse-practitioners are nurses with advanced qualifications who often act much like primary-care physicians. Are they as effective as doctors at treating patients with chronic conditions? An experiment was conducted with 1316 patients who had been diagnosed with asthma, diabetes, or high blood pressure. Within each condition, patients were randomly assigned to either a doctor or a nurse-practitioner. The response variables included measures of the patients' health and of their satisfaction with their medical care after 6 months.[44]

(a) Which are the blocks in this experiment: the different diagnoses (asthma, and so on) or the type of care (nurse or doctor)? Why?

(b) Explain why a randomized block design is preferable to a completely randomized design here.

76. **Comparing cancer treatments** The progress of a type of cancer differs in women and men. Researchers want to design an experiment to compare three therapies for this cancer. They recruit 500 male and 300 female patients who are willing to serve as subjects.

(a) Which are the blocks in this experiment: the cancer therapies or the two sexes? Why?

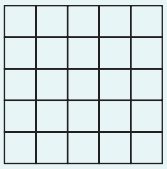
(b) What are the advantages of a randomized block design over a completely randomized design using these 800 subjects?

(c) Suppose the researchers had 800 male and no female subjects available for the study. What advantage would this offer? What disadvantage?

77. **In the cornfield** An agriculture researcher wants to compare the yield of 5 corn varieties: A, B, C, D, and E. The field in which the experiment will be carried out increases in fertility from north to south. The researcher therefore divides the field into 25 plots of equal size, arranged in 5 east–west rows of 5 plots each, as shown in the diagram.

pg 254

North



(a) Explain why a randomized block design would be better than a completely randomized design in this setting.

(b) Should the researcher use the rows or the columns of the field as blocks? Justify your answer.

(c) Use technology or Table D to carry out the random assignment required by your design. Explain your method clearly.

78. **Comparing weight-loss treatments** Twenty overweight females have agreed to participate in a study of the effectiveness of four weight-loss treatments: A, B, C, and D. The researcher first calculates how overweight each subject is by comparing the subject's actual weight with her "ideal" weight. The subjects and their excess weights in pounds are as follows:

Birnbaum	35	Hernandez	25	Moses	25	Smith	29
Brown	34	Jackson	33	Nevesky	39	Stall	33
Brunk	30	Kendall	28	Obrach	30	Tran	35
Cruz	34	Loren	32	Rodriguez	30	Wilansky	42
Deng	24	Mann	28	Santiago	27	Williams	22

The response variable is the weight lost after 8 weeks of treatment. Previous studies have shown that the effects of a diet may vary based on a subject's initial weight.

(a) Explain why a randomized block design would be better than a completely randomized design in this setting.

(b) Should researchers form blocks of size 4 based on subjects' last names in alphabetical order or by how overweight the subjects are? Explain.

(c) Use technology or Table D to carry out the random assignment required by your design. Explain your method clearly.

79. **Aw, rats!** A nutrition experimenter intends to compare the weight gain of newly weaned male rats fed Diet A with that of rats fed Diet B. To do this, she will feed each diet to 10 rats. She has available 10 rats from one litter and 10 rats from a second litter. Rats in the first litter appear to be slightly healthier.

(a) If the 10 rats from Litter 1 were fed Diet A, the effects of genetics and diet would be confounded, and the experiment would be biased. Explain this statement carefully.

(b) Describe a better design for this experiment.

80. **Technology for teaching statistics** The Brigham Young University (BYU) statistics department is performing experiments to compare teaching methods. Response variables include students' final-exam scores and a measure of their attitude toward statistics. One study compares two levels of technology for large lectures: standard (overhead projectors and chalk) and multimedia. There are eight lecture sections of a basic statistics course at BYU, each with about 200 students. There are four instructors, each of whom teaches two sections.[45] Suppose the sections and lecturers are as follows:

Section	Lecturer	Section	Lecturer
1	Hilton	5	Tolley
2	Christensen	6	Hilton
3	Hadfield	7	Tolley
4	Hadfield	8	Christensen

(a) Suppose we randomly assign two lecturers to use standard technology in their sections and the other two lecturers to use multimedia technology. Explain how this could lead to confounding.

(b) Describe a better design for this experiment.

81. **Look, Ma, no hands!** Does talking on a hands-free cell phone distract drivers? Researchers recruit 40 student subjects for an experiment to investigate this question. They have a driving simulator equipped with a hands-free phone for use in the study. Each subject will complete two sessions in the simulator: one while talking on the hands-free phone and the other while just driving. The order of the two sessions for each subject will be determined at random. The route, driving conditions, and traffic flow will be the same in both sessions.

(a) What type of design did the researchers use in their study?

(b) Explain why the researchers chose this design instead of a completely randomized design.

(c) Why is it important to randomly assign the order of the treatments?

(d) Explain how and why researchers controlled for other variables in this experiment.

82. **Chocolate gets my heart pumping** Cardiologists at Athens Medical School in Greece wanted to test whether chocolate affected blood flow in the blood vessels. The researchers recruited 17 healthy young volunteers, who were each given a 3.5-ounce bar of dark chocolate, either bittersweet or fake chocolate. On another day, the volunteers received the other treatment. The order in which subjects received the bittersweet and fake chocolate was determined at random. The subjects had no chocolate outside the study, and investigators didn't know whether a subject had eaten the real or the fake chocolate. An ultrasound was taken of each

volunteer's upper arm to see the functioning of the cells in the walls of the main artery. The researchers found that blood vessel function was improved when the subjects ate bittersweet chocolate, and that there were no such changes when they ate the placebo (fake chocolate).[46]

(a) What type of design did the researchers use in their study?

(b) Explain why the researchers chose this design instead of a completely randomized design.

(c) Why is it important to randomly assign the order of the treatments for the subjects?

(d) Explain how and why researchers controlled for other variables in this experiment.

83. **Room temperature and dexterity** An expert on worker performance is interested in the effect of room temperature on the performance of tasks requiring manual dexterity. She chooses temperatures of 70°F and 90°F as treatments. The response variable is the number of correct insertions, during a 30-minute period, in a peg-and-hole apparatus that requires the use of both hands simultaneously. Each subject is trained on the apparatus and then asked to make as many insertions as possible in 30 minutes of continuous effort.

(a) Describe a completely randomized design to compare dexterity at 70° and 90° using 20 volunteer subjects.

(b) Because individuals differ greatly in dexterity, the wide variation in individual scores may hide the systematic effect of temperature unless there are many subjects in each group. Describe in detail the design of a matched pairs experiment in which each subject serves as his or her own control.

84. **Carbon dioxide and tree growth** The concentration of carbon dioxide (CO_2) in the atmosphere is increasing rapidly due to our use of fossil fuels. Because plants use CO_2 to fuel photosynthesis, more CO_2 may cause trees and other plants to grow faster. An elaborate apparatus allows researchers to pipe extra CO_2 to a 30-meter circle of forest. We want to compare the growth in base area of trees in treated and untreated areas to see if extra CO_2 does in fact increase growth. We can afford to treat three circular areas.[47]

(a) Describe the design of a completely randomized experiment using six well-separated 30-meter circular areas in a pine forest. Sketch the circles and carry out the randomization your design calls for.

(b) Areas within the forest may differ in soil fertility. Describe a matched pairs design using three pairs of circles that will account for the extra variation due to different fertility. Sketch the circles and carry out the randomization your design calls for.

85. **Got deodorant?** A group of students wants to perform an experiment to determine whether Brand A or Brand B deodorant lasts longer. One group member suggests the following design: Recruit 40 student volunteers—20 male and 20 female. Separate by gender, because male and female bodies might respond differently to deodorant. Give all the males Brand A deodorant and all the females Brand B. Have each student rate how well the deodorant is still working at the end of the school day on a 0 to 10 scale. Then compare ratings for the two treatments.

(a) Identify any flaws you see in the proposed design for this experiment.

(b) Describe how you would design the experiment. Explain how your design addresses each of the problems you identified in part (a).

86. **Close shave** Which of two brands (X or Y) of electric razor shaves closer? Researchers want to design and carry out an experiment to answer this question using 50 adult male volunteers. Here's one idea: Have all 50 subjects shave the left sides of their faces with the Brand X razor and shave the right sides of their faces with the Brand Y razor. Then have each man decide which razor gave the closer shave and compile the results.

(a) Identify any flaws you see in the proposed design for this experiment.

(b) Describe how you would design the experiment. Explain how your design addresses each of the problems you identified in part (a).

Multiple choice: Select the best answer for Exercises 87 to 94.

87. Can changing diet reduce high blood pressure? Vegetarian diets and low-salt diets are both promising. Men with high blood pressure are assigned at random to four diets: (1) normal diet with unrestricted salt; (2) vegetarian with unrestricted salt; (3) normal with restricted salt; and (4) vegetarian with restricted salt. This experiment has

(a) one factor, the type of diet.

(b) two factors, high blood pressure and type of diet.

(c) two factors, normal/vegetarian diet and unrestricted/restricted salt.

(d) three factors, men, high blood pressure, and type of diet.

(e) four factors, the four diets being compared.

88. In the experiment of the previous exercise, the subjects were randomly assigned to the different treatments. What is the most important reason for this random assignment?

(a) Random assignment eliminates the effects of other variables such as stress and body weight.

(b) Random assignment is a good way to create groups of subjects that are roughly equivalent at the beginning of the experiment.

(c) Random assignment makes it possible to make a conclusion about all men.

(d) Random assignment reduces the amount of variation in blood pressure.

(e) Random assignment prevents the placebo effect from ruining the results of the study.

89. To investigate whether standing up while studying affects performance in an algebra class, a teacher assigns half of the 30 students in his class to stand up while studying and assigns the other half to not stand up while studying. To determine who receives which treatment, the teacher identifies the two students who did best on the last exam and randomly assigns one to stand and one to not stand. The teacher does the same for the next two highest-scoring students and continues in this manner until each student is assigned a treatment. Which of the following best describes this plan?

(a) This is an observational study.
(b) This is an experiment with blocking.
(c) This is a completely randomized experiment.
(d) This is a stratified random sample.
(e) This is a cluster sample.

90. A gardener wants to try different combinations of fertilizer (none, 1 cup, 2 cups) and mulch (none, wood chips, pine needles, plastic) to determine which combination produces the highest yield for a variety of green beans. He has 60 green-bean plants to use in the experiment. If he wants an equal number of plants to be assigned to each treatment, how many plants will be assigned to each treatment?

(a) 1 (b) 3 (c) 4 (d) 5 (e) 12

91. Corn variety 1 yielded 140 bushels per acre last year at a research farm. This year, corn variety 2, planted in the same location, yielded only 110 bushels per acre. Based on these results, is it reasonable to conclude that corn variety 1 is more productive than corn variety 2?

(a) Yes, because 140 bushels per acre is greater than 110 bushels per acre.
(b) Yes, because the study was done at a research farm.
(c) No, because there may be other differences between the two years besides the corn variety.
(d) No, because there was no use of a placebo in the experiment.
(e) No, because the experiment wasn't double-blind.

92. A report in a medical journal notes that the risk of developing Alzheimer's disease among subjects who regularly opted to take the drug ibuprofen was about half the risk among those who did not. Is this good evidence that ibuprofen is effective in preventing Alzheimer's disease?

(a) Yes, because the study was a randomized, comparative experiment.
(b) No, because the effect of ibuprofen is confounded with the placebo effect.
(c) Yes, because the results were published in a reputable professional journal.
(d) No, because this is an observational study. An experiment would be needed to confirm (or not confirm) the observed effect.
(e) Yes, because a 50% reduction can't happen just by chance.

93. A farmer is conducting an experiment to determine which variety of apple tree, Fuji or Gala, will produce more fruit in his orchard. The orchard is divided into 20 equally sized square plots. He has 10 trees of each variety and randomly assigns each tree to a separate plot in the orchard. What are the experimental unit(s) in this study?

(a) The trees (c) The apples (e) The orchard
(b) The plots (d) The farmer

94. Two essential features of all statistically designed experiments are

(a) compare several treatments; use the double-blind method.
(b) compare several treatments; use chance to assign subjects to treatments.
(c) always have a placebo group; use the double-blind method.
(d) use a block design; use chance to assign subjects to treatments.
(e) use enough subjects; always have a control group.

95. **Seed weights** (2.2) Biological measurements on the same species often follow a Normal distribution quite closely. The weights of seeds of a variety of winged bean are approximately Normal with mean 525 milligrams (mg) and standard deviation 110 mg.

(a) What percent of seeds weigh more than 500 mg? Show your method.
(b) If we discard the lightest 10% of these seeds, what is the smallest weight among the remaining seeds? Show your method.

96. **Twins** (1.3, 3.1) A researcher studied a group of identical twins who had been separated and adopted at birth. In each case, one twin (Twin A) was adopted by a low-income family and the other (Twin B) by a high-income family. Both twins were given an IQ test as adults. Here are their scores:[48]

Twin A:	120	99	99	94	111	97	99	94	104	114	113	100
Twin B:	128	104	108	100	116	105	100	100	103	124	114	112

(a) How well does one twin's IQ predict the other's? Give appropriate evidence to support your answer.
(b) Do identical twins living in low-income homes tend to have lower IQs later in life than their twins who live in high-income homes? Give appropriate evidence to support your answer.

4.3 Using Studies Wisely

WHAT YOU WILL LEARN **By the end of the section, you should be able to:**

- Describe the scope of inference that is appropriate in a statistical study.
- Evaluate whether a statistical study has been carried out in an ethical manner.*

Researchers who conduct statistical studies often want to draw conclusions (make inferences) that go beyond the data they produce. Here are two examples.

- The U.S. Census Bureau carries out a monthly Current Population Survey of about 60,000 households. Their goal is to use data from these randomly selected households to estimate the percent of unemployed individuals in the population.
- Scientists performed an experiment that randomly assigned 21 volunteer subjects to one of two treatments: sleep deprivation for one night or unrestricted sleep. The experimenters hoped to show that sleep deprivation causes a decrease in performance two days later.[49]

What type of inference can be made from a particular study? The answer depends on the design of the study.

Scope of Inference

In the Census Bureau's sample survey, the individuals who responded were chosen at random from the population of interest. Random sampling avoids bias and produces trustworthy estimates of the truth about the population. The Census Bureau should be safe making an *inference about the population* based on the results of the sample.

In the sleep deprivation experiment, subjects were randomly assigned to the sleep deprivation and unrestricted sleep treatments. Random assignment helps ensure that the two groups of subjects are as alike as possible before the treatments are imposed. If the unrestricted sleep group performs much better than the sleep deprivation group, and the difference is too large to be explained by chance variation in the random assignment, it must be due to the treatments. In that case, the scientists could safely conclude that sleep deprivation caused the decrease in performance. That is, they can make an *inference about cause and effect*. However, because the experiment used volunteer subjects, this limits scientists' ability to generalize their findings to some larger population of individuals.

Let's recap what we've learned about the scope of inference in a statistical study. Random selection of individuals allows inference about the population. Random assignment of individuals to groups permits inference about cause and effect. The following chart summarizes the possibilities.[50]

*This is an important topic, but it is not required for the AP® Statistics exam.

Both random sampling and random assignment introduce chance variation into a statistical study. When performing inference, statisticians use the laws of probability to describe this chance variation. You'll learn how this works later in the book.

		Were individuals randomly assigned to groups?	
		Yes	**No**
Were individuals randomly selected?	**Yes**	Inference about the population: YES Inference about cause and effect: YES	Inference about the population: YES Inference about cause and effect: NO
	No	Inference about the population: NO Inference about cause and effect: YES	Inference about the population: NO Inference about cause and effect: NO

Well-designed experiments randomly assign individuals to treatment groups. However, most experiments don't select experimental units at random from the larger population. That limits such experiments to inference about cause and effect. Observational studies don't randomly assign individuals to groups, which rules out inference about cause and effect. An observational study that uses random sampling can make an inference about the population. The following example illustrates all four cases from the table above in a single setting.

EXAMPLE Vitamin C and Canker Sores

Determining scope of inference

A small-town dentist wants to know if a daily dose of 500 milligrams (mg) of vitamin C will result in fewer canker sores in the mouth than taking no vitamin C.[51]

The dentist is considering the following four study designs:

Design 1: Get all dental patients in town with appointments in the next two weeks to take part in a study. Give each patient a survey with two questions: (1) Do you take at least 500 mg of vitamin C each day? (2) Do you frequently have canker sores? Based on patients' answers to Question 1, divide them into two groups: those who take at least 500 mg of vitamin C daily and those who don't.

Design 2: Get all dental patients in town with appointments in the next two weeks to take part in a study. Randomly assign half of them to take 500 mg of vitamin C each day and the other half to abstain from taking vitamin C for three months.

Design 3: Select a random sample of dental patients in town and get them to take part in a study. Divide the patients into two groups as in Design 1.

Design 4: Select a random sample of dental patients in town and get them to take part in a study. Randomly assign half of them to take 500 mg of vitamin C each day and the other half to abstain from taking vitamin C for three months.

For whichever design the dentist chooses, suppose she compares the proportion of patients in each group who complain of canker sores. Also suppose that she finds a statistically significant difference, with a smaller proportion of those taking vitamin C having canker sores.

PROBLEM: What can the dentist conclude for each design?

SOLUTION:

Design 1: Because the patients were not randomly selected, the dentist cannot infer that this result holds for a larger population of dental patients. This was an observational study because no treatments were deliberately imposed on the patients. With no random assignment to the two

groups, no inference about cause and effect can be made. The dentist just knows that for these patients, those who took vitamin C had fewer canker sores than those who didn't.

Design 2: As in Design 1, the dentist can't make any inference about this result holding for a larger population of dental patients. However, the treatments were randomly assigned to the subjects. Assuming proper control in the experiment, she can conclude that taking vitamin C reduced the chance of getting canker sores in her subjects.

Design 3: Because the patients were randomly selected from the population of dental patients in the town, the dentist can generalize the results of this study to the population. Because this was an observational study, no inference about cause and effect can be made. The dentist would conclude that for the population of dental patients in this town, those taking vitamin C have fewer canker sores than those who don't. She can't say whether the vitamin C causes this reduction or some other confounding variable.

Design 4: As in Design 3, the random sampling allows the dentist to generalize the results of this study to the population of dental patients in the town. As in Design 2, the random assignment would allow her to conclude (assuming proper control in the experiment) that taking vitamin C reduced the chance of getting canker sores. So the dentist would conclude that for the population of dental patients in this town, those taking vitamin C will tend to have fewer canker sores than those who don't due to the vitamin C.

For Practice *Try Exercise* 101

The Challenges of Establishing Causation

A well-designed experiment tells us that changes in the explanatory variable cause changes in the response variable. More precisely, it tells us that this happened for specific individuals in the specific environment of this specific experiment. A serious threat is that the treatments, the subjects, or the environment of our experiment may not be realistic. *Lack of realism* can limit our ability to apply the conclusions of an experiment to the settings of greatest interest.

EXAMPLE

Do Center Brake Lights Reduce Rear-End Crashes?

Lack of realism

Do those high center brake lights, required on all cars sold in the United States since 1986, really reduce rear-end collisions? Randomized comparative experiments with fleets of rental and business cars, done before the lights were required, showed that the third brake light reduced rear-end collisions by as much as 50%. But requiring the third light in all cars led to only a 5% drop.

What happened? Most cars did not have the extra brake light when the experiments were carried out, so it caught the eye of following drivers. Now that almost all cars have the third light, they no longer capture attention.

In some cases, it isn't practical or even ethical to do an experiment. Consider these important questions:

- Does texting while driving increase the risk of having an accident?
- Does going to church regularly help people live longer?
- Does smoking cause lung cancer?

To answer these cause-and-effect questions, we just need to perform a randomized comparative experiment. Unfortunately, we can't randomly assign people to text while driving or to attend church or to smoke cigarettes. The best data we have about these and many other cause-and-effect questions come from observational studies.

It is sometimes possible to build a strong case for causation in the absence of experiments. The evidence that smoking causes lung cancer is about as strong as nonexperimental evidence can be.

EXAMPLE

Does Smoking Cause Lung Cancer?

Living with observational studies

Doctors had long observed that most lung cancer patients were smokers. Comparison of smokers and similar nonsmokers showed a very strong association between smoking and death from lung cancer. Could the association be due to some other variable? Is there some genetic factor that makes people both more likely to get addicted to nicotine and to develop lung cancer? If so, then smoking and lung cancer would be strongly associated even if smoking had no direct effect on the lungs. Or maybe confounding is to blame. It might be that smokers live unhealthy lives in other ways (diet, alcohol, lack of exercise) and that some other habit confounded with smoking is a cause of lung cancer. How were these objections overcome?

What are the criteria for establishing causation when we can't do an experiment?

- *The association is strong.* The association between smoking and lung cancer is very strong.
- *The association is consistent.* Many studies of different kinds of people in many countries link smoking to lung cancer. That reduces the chance that some other variable specific to one group or one study explains the association.
- *Larger values of the explanatory variable are associated with stronger responses.* People who smoke more cigarettes per day or who smoke over a longer period get lung cancer more often. People who stop smoking reduce their risk.
- *The alleged cause precedes the effect in time.* Lung cancer develops after years of smoking. The number of men dying of lung cancer rose as smoking became more common, with a lag of about 30 years. Lung cancer kills more men than any other form of cancer. Lung cancer was rare among women until women began to smoke. Lung cancer in women rose along with smoking, again with a lag of about 30 years, and has now passed breast cancer as the leading cause of cancer death among women.
- *The alleged cause is plausible.* Experiments with animals show that tars from cigarette smoke do cause cancer.

Medical authorities do not hesitate to say that smoking causes lung cancer. The U.S. Surgeon General states that cigarette smoking is "the largest avoidable

cause of death and disability in the United States."[53] The evidence for causation is overwhelming—but it is not as strong as the evidence provided by well-designed experiments. Conducting an experiment in which some subjects were forced to smoke and others were not allowed to would be unethical. In cases like this, observational studies are our best source of reliable information.

Data Ethics*

Medical professionals are taught to follow the basic principle "First, do no harm." Shouldn't those who carry out statistical studies follow the same principle? Most reasonable people think so. But this may not always be as simple as it sounds. Decide whether you think each of the following studies is ethical or unethical:

- A promising new drug has been developed for treating cancer in humans. Before giving the drug to human subjects, researchers want to administer the drug to animals to see if there are any potentially serious side effects.
- Are companies discriminating against some individuals in the hiring process? To find out, researchers prepare several equivalent résumés for fictitious job applicants, with the only difference being the gender of the applicant. They send the fake résumés to companies advertising positions and keep track of the number of males and females who are contacted for interviews.
- Will people try to stop someone from driving drunk? A television news program hires an actor to play a drunk driver and uses a hidden camera to record the behavior of individuals who encounter the driver.

The most complex issues of data ethics arise when we collect data from people. The ethical difficulties are more severe for experiments that impose some treatment on people than for sample surveys that simply gather information. Trials of new medical treatments, for example, can do harm as well as good to their subjects. Here are some basic standards of data ethics that must be obeyed by all studies that gather data from human subjects, both observational studies and experiments.

BASIC DATA ETHICS

All planned studies must be reviewed in advance by an **institutional review board** charged with protecting the safety and well-being of the subjects.

All individuals who are subjects in a study must give their **informed consent** before data are collected.

All individual data must be kept **confidential**. Only statistical summaries for groups of subjects may be made public.

The law requires that studies carried out or funded by the federal government obey these principles.[54] But neither the law nor the consensus of experts is completely clear about the details of their application.

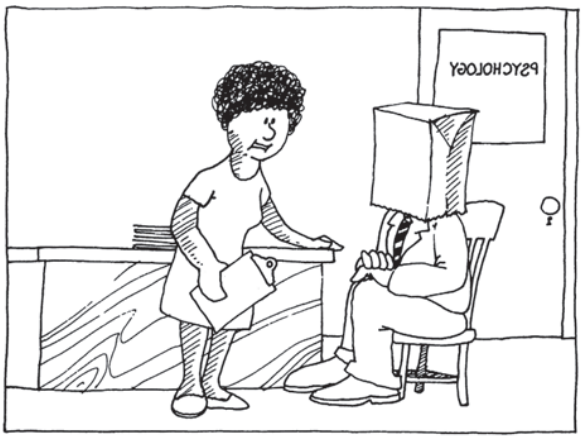
Institutional review boards The purpose of an institutional review board is not to decide whether a proposed study will produce valuable information or whether it is statistically sound. The board's purpose is, in the words of one uni-

*This is an important topic, but it is not required for the AP® Statistics exam.

versity's board, "to protect the rights and welfare of human subjects (including patients) recruited to participate in research activities." The board reviews the plan of the study and can require changes. It reviews the consent form to be sure that subjects are informed about the nature of the study and about any potential risks. Once research begins, the board monitors its progress at least once a year.

Informed consent Both words in the phrase "informed consent" are important, and both can be controversial. Subjects must be *informed* in advance about the nature of a study and any risk of harm it may bring. In the case of a sample survey, physical harm is not possible. But a survey on sensitive issues could result in emotional harm. The participants should be told what kinds of questions the survey will ask and about how much of their time it will take. Experimenters must tell subjects the nature and purpose of the study and outline possible risks. Subjects must then *consent* in writing.

Confidentiality Ethical problems do not disappear once a study has been cleared by the review board, has obtained consent from its participants, and has actually collected data about them. It is important to protect individuals' privacy by keeping all data about them **confidential.** The report of an opinion poll may say what percent of the 1200 respondents believed that legal immigration should be reduced. It may not report what *you* said about this or any other issue.



"I realize the participants in this study are to be anonymous, but you're going to have to expose your eyes."

Confidentiality is not the same as **anonymity.** Anonymity means that individuals are anonymous—their names are not known even to the director of the study. Anonymity is rare in statistical studies. Even where anonymity is possible (mainly in surveys conducted by mail), it prevents any follow-up to improve nonresponse or inform individuals of results.

Any breach of confidentiality is a serious violation of data ethics. The best practice is to separate the identity of the study's participants from the rest of the data at once. A clever computer search of several data bases might be able, by combining information, to identify you and learn a great deal about you even if your name and other identification have been removed from the data available for search. Privacy and confidentiality of data are hot issues among statisticians in the computer age.

ACTIVITY | Response bias

In this Activity, your team will design and conduct an experiment to investigate the effects of response bias in surveys.[52] You may choose the topic for your surveys, but you must design your experiment so that it can answer at least one of the following questions:

- Can the wording of a question create response bias?
- Do the characteristics of the interviewer create response bias?
- Does anonymity change the responses to sensitive questions?
- Does manipulating the answer choices change the response?

1. Write a proposal describing the design of your experiment. Be sure to include
   (a) your chosen topic and which of the above questions you'll try to answer.
   (b) a detailed description of how you will obtain your subjects (minimum of 50). Your plan must be practical!
   (c) what your questions will be and how they will be asked.
   (d) a clear explanation of how you will implement your design.
   (e) precautions you will take to collect data ethically.

   Here are two examples of successful student experiments:

"Make-Up," by Caryn S. and Trisha T. (all questions asked to males)
   i. "Do you find females who wear makeup attractive?" (questioner wearing makeup: 75% answered yes)
   ii. "Do you find females who wear makeup attractive?" (questioner not wearing makeup: 30% answered yes)

"Cartoons" by Sean W. and Brian H.
   i. "Do you watch cartoons?" (90% answered yes)
   ii. "Do you *still* watch cartoons?" (60% answered yes)

2. Once your teacher has approved your design, carry out the experiment. Record your data in a table.

3. Analyze your data. What conclusion do you draw? Provide appropriate graphical and numerical evidence to support your answer.

4. Prepare a report that includes the data you collected, your analysis from Step 3, and a discussion of any problems you encountered and how you dealt with them.

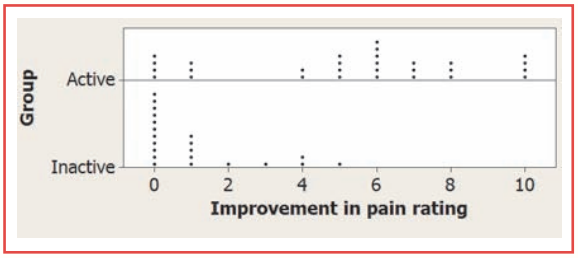
case closed

Can Magnets Help Reduce Pain?

Re-read the chapter-opening Case Study on page 207. Then use what you have learned in this chapter to help answer the following questions.

1. Why is the magnet study an experiment, not an observational study?
2. What type of design was used in this experiment? Identify the experimental units, the treatments, and the response variable.
3. There were two distinct purposes for having the doctors select a sealed envelope at random from the box. Describe them both.

4. The dotplot shows the improvement in pain ratings for both groups. Write a few sentences comparing the two distributions.



5. The mean difference in pain ratings was 5.24 for the active-magnet group and 1.10 for the inactive-magnet group. This difference is statistically significant. What conclusion should we draw?

Section 4.3 Summary

- Most statistical studies aim to make inferences that go beyond the data actually produced. **Inference about a population** requires that the individuals taking part in a study be randomly selected from the population. A well-designed experiment that randomly assigns experimental units to treatments allows **inference about cause and effect**.
- **Lack of realism** in an experiment can prevent us from generalizing its results.
- In the absence of an experiment, good evidence of causation requires a strong association that appears consistently in many studies, a clear explanation for the alleged causal link, and careful examination of other variables.
- Studies involving humans must be screened in advance by an **institutional review board**. All participants must give their **informed consent** before taking part. Any information about the individuals in the study must be kept **confidential**.

Section 4.3 Exercises

97. **Random sampling versus random assignment** Explain the difference between the types of inference that can be made as a result of random sampling and random assignment.

98. **Observation versus experimentation** Explain the difference between the types of inference than can usually be made from an observational study and an experiment.

99. **Foster care versus orphanages** Do abandoned children placed in foster homes do better than similar children placed in an institution? The Bucharest Early Intervention Project found that the answer is a clear "Yes." The subjects were 136 young children abandoned at birth and living in orphanages in Bucharest, Romania. Half of the children, chosen at random, were placed in foster homes. The other half remained in the orphanages.[55] (Foster care was not easily available in Romania at the time and so was paid for by the study.) What conclusion can we draw from this study? Explain.

100. **Frozen batteries** Will storing batteries in a freezer make them last longer? To find out, a company that

produces batteries takes a random sample of 100 AA batteries from its warehouse. The company statistician randomly assigns 50 batteries to be stored in the freezer and the other 50 to be stored at room temperature for 3 years. At the end of that time period, each battery's charge is tested. *Result:* Batteries stored in the freezer had a higher average charge, and the difference between the groups was statistically significant. What conclusion can we draw from this study? Explain.

101. **Who talks more—women or men?** According to Louann Brizendine, author of *The Female Brain*, women say nearly three times as many words per day as men. Skeptical researchers devised a study to test this claim. They used electronic devices to record the talking patterns of 396 university students who volunteered to participate in the study. The device was programmed to record 30 seconds of sound every 12.5 minutes without the carrier's knowledge. According to a published report of the study in *Scientific American*, "Men showed a slightly wider variability in words uttered.... But in the end, the sexes came out just about even in the daily averages: women at 16,215 words and men at 15,669."[56] This difference was not statistically significant. What conclusion can we draw from this study? Explain.

pg 267

102. **Attend church, live longer?** One of the better studies of the effect of regular attendance at religious services gathered data from a random sample of 3617 adults. The researchers then measured lots of variables, not just the explanatory variable (religious activities) and the response variable (length of life). A news article said: "Churchgoers were more likely to be nonsmokers, physically active, and at their right weight. But even after health behaviors were taken into account, those not attending religious services regularly still were about 25% more likely to have died."[57] What conclusion can we draw from this study? Explain.

103. **Daytime running lights** Canada and the European Union require that cars be equipped with "daytime running lights," headlights that automatically come on at a low level when the car is started. Many manufacturers are now equipping cars sold in the United States with running lights. Will running lights reduce accidents by making cars more visible? An experiment conducted in a driving simulator suggests that the answer may be "Yes." What concerns would you have about generalizing the results of such an experiment?

104. **Studying frustration** A psychologist wants to study the effects of failure and frustration on the relationships among members of a work team. She forms a team of students, brings them to the psychology lab, and has them play a game that requires teamwork. The game is rigged so that they lose regularly. The psychologist observes the students through a one-way window and notes the changes in their behavior during an evening of game playing. Can the psychologist generalize the results of her study to a team of employees that spends months developing a new product that never works right and is finally abandoned by their company? Explain.

105.* **Minimal risk?** You have been invited to serve on a college's institutional review board. You must decide whether several research proposals qualify for lighter review because they involve only minimal risk to subjects. Federal regulations say that "minimal risk" means the risks are no greater than "those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests." That's vague. Which of these do you think qualifies as "minimal risk"?

(a) Draw a drop of blood by pricking a finger to measure blood sugar.

(b) Draw blood from the arm for a full set of blood tests.

(c) Insert a tube that remains in the arm, so that blood can be drawn regularly.

106.* **Who reviews?** Government regulations require that institutional review boards consist of at least five people, including at least one scientist, one nonscientist, and one person from outside the institution. Most boards are larger, but many contain just one outsider.

(a) Why should review boards contain people who are not scientists?

(b) Do you think that one outside member is enough? How would you choose that member? (For example, would you prefer a medical doctor? A member of the clergy? An activist for patients' rights?)

107.* **No consent needed?** In which of the circumstances below would you allow collecting personal information without the subjects' consent?

(a) A government agency takes a random sample of income tax returns to obtain information on the average income of people in different occupations. Only the incomes and occupations are recorded from the returns, not the names.

(b) A social psychologist attends public meetings of a religious group to study the behavior patterns of members.

(c) A social psychologist pretends to be converted to membership in a religious group and attends private meetings to study the behavior patterns of members.

108.* **Surveys of youth** A survey asked teenagers whether they had ever consumed an alcoholic beverage. Those who said "Yes" were then asked, "How old were you when you first consumed an alcoholic beverage?" Should consent of parents be required to ask minors about alcohol, drugs, and other such issues, or is consent of the minors themselves enough? Give reasons for your opinion.

109.* **Anonymous? Confidential?** One of the most important nongovernment surveys in the United States is the National Opinion Research Center's General Social

*Exercises 105 to 112: This is an important topic, but it is not required for the AP® Statistics exam.

Survey. The GSS regularly monitors public opinion on a wide variety of political and social issues. Interviews are conducted in person in the subject's home. Are a subject's responses to GSS questions anonymous, confidential, or both? Explain your answer.

110.* **Anonymous? Confidential?** Texas A&M, like many universities, offers screening for HIV, the virus that causes AIDS. Students may choose either anonymous or confidential screening. An announcement says, "Persons who sign up for screening will be assigned a number so that they do not have to give their name." They can learn the results of the test by telephone, still without giving their name. Does this describe the *anonymous* or the *confidential* screening? Why?

111.* **The Willowbrook hepatitis studies** In the 1960s, children entering the Willowbrook State School, an institution for the intellectually disabled on Staten Island in New York, were deliberately infected with hepatitis. The researchers argued that almost all children in the institution quickly became infected anyway. The studies showed for the first time that two strains of hepatitis existed. This finding contributed to the development of effective vaccines. Despite these valuable results, the Willowbrook studies are now considered an example of unethical research. Explain why, according to current ethical standards, useful results are not enough to allow a study.

112.* **Unequal benefits** Researchers on aging proposed to investigate the effect of supplemental health services on the quality of life of older people. Eligible patients on the rolls of a large medical clinic were to be randomly assigned to treatment and control groups. The treatment group would be offered hearing aids, dentures, transportation, and other services not available without charge to the control group. The review board felt that providing these services to some but not other persons in the same institution raised ethical questions. Do you agree?

113. **Animal testing** (1.1) "It is right to use animals for medical testing if it might save human lives." The General Social Survey asked 1152 adults to react to this statement. Here is the two-way table of their responses:

	Male	Female
Strongly agree	76	59
Agree	270	247
Neither agree nor disagree	87	139
Disagree	61	123
Strongly disagree	22	68

How do the distributions of opinion differ between men and women? Give appropriate graphical and numerical evidence to support your answer.

114. **Initial public offerings** (1.3) The business magazine *Forbes* reports that 4567 companies sold their first stock to the public between 1990 and 2000. The *mean* change in the stock price of these companies since the first stock was issued was +111%. The *median* change was −31%.[58] Explain how this could happen.

*Exercises 105 to 112: This is an important topic, but it is not required for the AP® Statistics exam.

FRAPPY! Free Response AP® Problem, Yay!

The following problem is modeled after actual AP® Statistics exam free response questions. Your task is to generate a complete, concise response in 15 minutes.

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

In a recent study, 166 adults from the St. Louis area were recruited and randomly assigned to receive one of two treatments for a sinus infection. Half of the subjects received an antibiotic (amoxicillin) and the other half received a placebo.[59]

(a) Describe how the researchers could have assigned treatments to subjects if they wanted to use a completely randomized design.

(b) All the subjects in the experiment had moderate, severe, or very severe symptoms at the beginning of the study. Describe one statistical benefit and one statistical drawback for using subjects with moderate, severe, or very severe symptoms instead of just using subjects with very severe symptoms.

(c) At different stages during the next month, all subjects took the sino-nasal outcome test. After 10 days, the difference in average test scores was *not* statistically significant. In this context, explain what it means for the difference to be not statistically significant.

(d) One possible way that researchers could have improved the study is to use a randomized block design. Explain how the researchers could have incorporated blocking in their design.

After you finish, you can view two example solutions on the book's Web site (www.whfreeman.com/tps5e). Determine whether you think each solution is "complete," "substantial," "developing," or "minimal." If the solution is not complete, what improvements would you suggest to the student who wrote it? Finally, your teacher will provide you with a scoring rubric. Score your response and note what, if anything, you would do differently to improve your own score.

Chapter Review

Section 4.1: Sampling and Surveys

In this section, you learned that a population is the group of all individuals that we want information about. A sample is the subset of the population that we use to gather this information. The goal of most sample surveys is to use sample information to draw conclusions about the population. Choosing people for a sample because they are located nearby or letting people choose whether or not to be in the sample are poor ways to choose a sample. Because convenience samples and voluntary response samples will produce estimates that are consistently too large or consistently too small, these methods of choosing a sample are biased.

To avoid bias in the way the sample is formed, the members of the sample should be chosen at random. One way to do this is with a simple random sample (SRS), which is equivalent to pulling well-mixed slips of paper from a hat. It is often more convenient to select an SRS using technology or a table of random digits.

Two other random sampling methods are stratified sampling and cluster sampling. To obtain a stratified random sample, divide the population into groups (strata) of similar individuals, take an SRS from each stratum, and combine the chosen individuals to form the sample. Stratified random samples can produce estimates with much greater precision than simple random samples. To obtain a cluster sample, divide the population into groups (clusters) of individuals that are in similar locations, randomly select clusters, and use every individual in the chosen clusters. Cluster samples are easier to obtain than simple random samples or stratified random samples, but they may not produce very precise estimates.

Finally, you learned about other issues in sample surveys that can lead to bias: undercoverage occurs when the sampling method systematically excludes one part of the population. Nonresponse describes when answers cannot be obtained from some people that were chosen to be in the sample. Bias can also result when some people in the sample don't give accurate responses due to question wording, interviewer characteristics, or other factors.

Section 4.2: Experiments

In this section, you learned about the difference between observational studies and experiments. Experiments deliberately impose a treatment to see if there is a cause-and-effect relationship between two variables. Observational studies look at relationships between two variables, but cannot show cause and effect because other variables may be confounded with the explanatory variable. Variables are confounded when it is impossible to determine which of the variables is causing a change in the response variable.

A common type of experiment uses a completely randomized design. In this type of design, the experimental units are divided into groups, one group for each of the treatments. To determine which experimental units are in which group, we use random assignment. With random assignment, the effects of variables (other than the explanatory variable) are roughly balanced out between the groups. Replication means giving each treatment to as many experimental units as possible. This makes it easier to see the effects of the treatments because the effects of other variables are more likely to be balanced among the treatment groups.

During an experiment, it is important that other variables be controlled (kept the same) for each experimental unit. Doing so helps avoid confounding and removes a possible source of variation in the response variable. Also, beware of the placebo effect—the tendency for people to improve because they expect to, not because of the treatment they are receiving. One way to make sure that all experimental units have the same expectations is to make them blind—unaware of which treatment they are receiving. When the person measuring the response variable is also blind, the experiment is called double-blind.

The results of an experiment are statistically significant if the difference in the response is too large to be accounted for by the random assignment of experimental units to treatments. To make it more likely to obtain statistically significant results, experiments can incorporate blocking. Blocking in experiments is similar to stratifying in sampling. To form blocks, group together experimental units that are similar with respect to a variable that is associated with the response. Then randomly assign the treatments within each block. A design that uses blocks with two experimental units is called a matched pairs design. Blocking helps us estimate the effects of the treatments more precisely because we can account for the variability introduced by the variables used to form the blocks.

Section 4.3: Using Studies Wisely

In this section, you learned that the different types of conclusions we can draw depend on how the data are produced. When samples are selected at random, we can make inferences about the population from which the sample was drawn. When treatments are applied to groups formed at random, we can conclude cause and effect.

Making a cause-and-effect conclusion is often difficult because it is impossible or unethical to perform certain types of experiments. Good data ethics requires that studies should be approved by an institutional review board, subjects should give informed consent, and individual data must be kept confidential.

What Did You Learn?

Learning Objective	Section	Related Example on Page(s)	Relevant Chapter Review Exercise(s)
Identify the population and sample in a statistical study.	4.1	210	R4.1
Identify voluntary response samples and convenience samples. Explain how these sampling methods can lead to bias.	4.1	213	R4.2
Describe how to obtain a random sample using slips of paper, technology, or a table of random digits.	4.1	214, 217	R4.3
Distinguish a simple random sample from a stratified random sample or cluster sample. Give the advantages and disadvantages of each sampling method.	4.1	221	R4.4
Explain how undercoverage, nonresponse, question wording, and other aspects of a sample survey can lead to bias.	4.1	226, 227	R4.5
Distinguish between an observational study and an experiment.	4.2	235	R4.6
Explain the concept of confounding and how it limits the ability to make cause-and-effect conclusions.	4.2	235	R4.6
Identify the experimental units, explanatory and response variables, and treatments in an experiment.	4.2	237, 239	R4.7
Explain the purpose of comparison, random assignment, control, and replication in an experiment.	4.2	243	R4.9
Describe a completely randomized design for an experiment, including how to randomly assign treatments using slips of paper, technology, or a table of random digits.	4.2	246	R4.7, R4.10
Describe the placebo effect and the purpose of blinding in an experiment.	4.2	247	R4.9
Interpret the meaning of statistically significant in the context of an experiment.	4.2	249 (Activity)	R4.9
Explain the purpose of blocking in an experiment. Describe a randomized block design or a matched pairs design for an experiment.	4.2	251, 254	R4.7, R4.10
Describe the scope of inference that is appropriate in a statistical study.	4.3	267	R4.8
*Evaluate whether a statistical study has been carried out in an ethical manner.	4.3	Discussion on 270	R4.11

*This is an important topic, but it is not required for the AP® Statistics exam.

Chapter 4 Chapter Review Exercises

These exercises are designed to help you review the important ideas and methods of the chapter.

R4.1 Ontario Health Survey The Ministry of Health in the province of Ontario, Canada, wants to know whether the national health care system is achieving its goals in the province. Much information about health care comes from patient records, but that source doesn't allow us to compare people who use health services with those who don't. So the Ministry of Health conducted the Ontario Health Survey, which interviewed a random sample of 61,239 people who live in the province of Ontario.[60]

(a) What is the population for this sample survey? What is the sample?

(b) The survey found that 76% of males and 86% of females in the sample had visited a general practitioner at least once in the past year. If a census were conducted, do you think that the percentages would be the same as in the sample? Explain.

R4.2 Bad sampling A large high school wants to gather student opinion about parking for students on campus. It isn't practical to contact all students.

(a) Give an example of a way to choose a voluntary response sample of students. Explain how this method could lead to bias.

(b) Give an example of a way to choose a convenience sample of students. Explain how this method could lead to bias.

R4.3 Drug testing A baseball team regularly conducts random drug tests on its players. The 25 members of the team are listed below.

Agarwal	Chen	Healy	Moser	Roberts
Andrews	Frank	Hixson	Musselman	Shen
Baer	Fuest	Lee	Pavnica	Smith
Berger	Fuhrmann	Lynch	Petrucelli	Sundheim
Brockman	Garcia	Milhalko	Reda	Wilson

(a) Explain how you would use the line of random digits below to select an SRS of 3 team members for a random drug test.

(b) Use your method from part (a) to choose the SRS using the digits below. Show your work.

17521 78009 46239 84569 03316

R4.4 Polling the faculty A researcher wants to study the attitudes of college faculty members about the work habits of entering freshmen. These attitudes appear to differ depending on the type of college. The American Association of University Professors classifies colleges as follows:

Class I: Offer doctorate degrees and award at least 15 per year.

Class IIA: Award degrees above the bachelor's but are not in Class I.

Class IIB: Award no degrees beyond the bachelor's.

Class III: Two-year colleges.

The researcher would like to survey about 200 faculty members. Would you recommend a simple random sample, stratified random sample, or cluster sample? Justify your answer.

R4.5 Been to the movies? An opinion poll calls 2000 ran domly chosen residential telephone numbers, then asks to speak with an adult member of the household. The interviewer asks, "How many movies have you watched in a movie theater in the past 12 months?" In all, 1131 people responded. The researchers used the responses to estimate the mean number of movies adults have watched in a movie theater in the past 12 months.

(a) Describe a potential source of bias related to the wording of the question. Suggest a change that would help fix this problem.

(b) Describe how using only residential phone numbers might lead to bias and how this will affect the estimate.

(c) Describe how nonresponse might lead to bias and how this will affect the estimate.

R4.6 Are anesthetics safe? The National Halothane Study was a major investigation of the safety of anesthetics used in surgery. Records of over 850,000 operations performed in 34 major hospitals showed the following death rates for four common anesthetics:[61]

Anesthetic:	A	B	C	D
Death rate:	1.7%	1.7%	3.4%	1.9%

There seems to be a clear association between the anesthetic used and the death rate of patients. Anesthetic C appears to be more dangerous.

(a) Explain why we call the National Halothane Study an observational study rather than an experiment, even though it compared the results of using different anesthetics in actual surgery.

(b) When the study looked at other variables that are related to a doctor's choice of anesthetic, it found that Anesthetic C was not causing extra deaths. Explain the concept of confounding in this context and identify a variable that might be confounded with the doctor's choice of anesthetic.

R4.7 **Ugly fries** Few people want to eat discolored french fries. Potatoes are kept refrigerated before being cut for french fries to prevent spoiling and preserve flavor. But immediate processing of cold potatoes causes discoloring due to complex chemical reactions. The potatoes must therefore be brought to room temperature before processing. Researchers want to design an experiment in which tasters will rate the color and flavor of french fries prepared from several groups of potatoes. The potatoes will be freshly picked or stored for a month at room temperature or stored for a month refrigerated. They will then be sliced and cooked either immediately or after an hour at room temperature.

(a) Identify the experimental units, the explanatory and response variables, and the treatments.

(b) The researchers plan to use a completely randomized design. Describe how they should assign treatments to the experimental units if there are 300 potatoes available for the experiment.

(c) The researchers decided to do a follow-up experiment using sweet potatoes as well as regular potatoes. Describe how they should change the design of the experiment to account for the addition of sweet potatoes.

R4.8 **Don't catch a cold!** A recent study of 1000 students at the University of Michigan investigated how to prevent catching the common cold. The students were randomly assigned to three different cold prevention methods for 6 weeks. Some wore masks, some wore masks and used hand sanitizer, and others took no precautions. The two groups who used masks reported 10–50% fewer cold symptoms than those who did not wear a mask.[62]

(a) Does this study allow for inference about a population? Explain.

(b) Does this study allow for inference about cause and effect? Explain.

R4.9 **An herb for depression?** Does the herb Saint-John's-wort relieve major depression? Here is an excerpt from the report of a study of this issue: "Design: Randomized, Double-Blind, Placebo-Controlled Clinical Trial."[63] The study concluded that the difference in effectiveness of Saint-John's-wort and a placebo was not statistically significant.

(a) How did the design of this experiment account for the placebo effect?

(b) Explain the purpose of the random assignment.

(c) Why is a double-blind design a good idea in this setting?

(d) Explain what "not statistically significant" means in this context.

R4.10 **How long did I work?** A psychologist wants to know if the difficulty of a task influences our estimate of how long we spend working at it. She designs two sets of mazes that subjects can work through on a computer. One set has easy mazes and the other has difficult mazes. Subjects work until told to stop (after 6 minutes, but subjects do not know this). They are then asked to estimate how long they worked. The psychologist has 30 students available to serve as subjects.

(a) Describe an experiment using a completely randomized design to learn the effect of difficulty on estimated time.

(b) Describe a matched pairs experimental design using the same 30 subjects.

(c) Which design would be more likely to detect a difference in the effects of the treatments? Explain.

R4.11* **Deceiving subjects** Students sign up to be subjects in a psychology experiment. When they arrive, they are told that interviews are running late and are taken to a waiting room. The experimenters then stage a theft of a valuable object left in the waiting room. Some subjects are alone with the thief, and others are in pairs—these are the treatments being compared. Will the subject report the theft?

(a) The students had agreed to take part in an unspecified study, and the true nature of the experiment is explained to them afterward. Does this meet the requirement of informed consent? Explain.

(b) What two other ethical principles should be followed in this study?

*This is an important topic, but it is not required for the AP® Statistics exam.

Chapter 4 AP® Statistics Practice Test

Section I: Multiple Choice *Select the best answer for each question.*

T4.1 When we take a census, we attempt to collect data from

(a) a stratified random sample.

(b) every individual chosen in a simple random sample.

(c) every individual in the population.

(d) a voluntary response sample.

(e) a convenience sample.

T4.2 You want to take a simple random sample (SRS) of 50 of the 816 students who live in a dormitory on

campus. You label the students 001 to 816 in alphabetical order. In the table of random digits, you read the entries

95592 94007 69769 33547 72450 16632 81194 14873

The first three students in your sample have labels

(a) 955, 929, 400.
(b) 400, 769, 769.
(c) 559, 294, 007.
(d) 929, 400, 769.
(e) 400, 769, 335.

T4.3 A study of treatments for angina (pain due to low blood supply to the heart) compared bypass surgery, angioplasty, and use of drugs. The study looked at the medical records of thousands of angina patients whose doctors had chosen one of these treatments. It found that the average survival time of patients given drugs was the highest. What do you conclude?

(a) This study proves that drugs prolong life and should be the treatment of choice.
(b) We can conclude that drugs prolong life because the study was a comparative experiment.
(c) We can't conclude that drugs prolong life because the patients were volunteers.
(d) We can't conclude that drugs prolong life because this was an observational study.
(e) We can't conclude that drugs prolong life because no placebo was used.

T4.4 Tonya wanted to estimate the average amount of time that students at her school spend on Facebook each day. She gets an alphabetical roster of students in the school from the registrar's office and numbers the students from 1 to 1137. Then Tonya uses a random number generator to pick 30 distinct labels from 1 to 1137. She surveys those 30 students about their Facebook use. Tonya's sample is a simple random sample because

(a) it was selected using a chance process.
(b) it gave every individual the same chance to be selected.
(c) it gave every possible sample of the same size an equal chance to be selected.
(d) it doesn't involve strata or clusters.
(e) it is guaranteed to be representative of the population.

T4.5 Consider an experiment to investigate the effectiveness of different insecticides in controlling pests and their impact on the productivity of tomato plants. What is the best reason for randomly assigning treatment levels (spraying or not spraying) to the experimental units (farms)?

(a) Random assignment allows researchers to generalize conclusions about the effectiveness of the insecticides to all farms.
(b) Random assignment will tend to average out all other uncontrolled factors such as soil fertility so that they are not confounded with the treatment effects.
(c) Random assignment eliminates the effects of other variables, like soil fertility.
(d) Random assignment eliminates chance variation in the responses.
(e) Random assignment helps avoid bias due to the placebo effect.

T4.6 The most important advantage of experiments over observational studies is that

(a) experiments are usually easier to carry out.
(b) experiments can give better evidence of causation.
(c) confounding cannot happen in experiments.
(d) an observational study cannot have a response variable.
(e) observational studies cannot use random samples.

T4.7 A TV station wishes to obtain information on the TV viewing habits in its market area. The market area contains one city of population 170,000, another city of 70,000, and four towns of about 5000 inhabitants each. The station suspects that the viewing habits may be different in larger and smaller cities and in the rural areas. Which of the following sampling designs would give the type of information that the station requires?

(a) A cluster sample using the cities and towns as clusters
(b) A convenience sample from the market area
(c) A simple random sample from the market area
(d) A stratified sample from the cities and towns in the market area
(e) An online poll that invites all people from the cities and towns in the market area to participate

T4.8 *Bias* in a sampling method is

(a) any difference between the sample result and the truth about the population.
(b) the difference between the sample result and the truth about the population due to using chance to select a sample.
(c) any difference between the sample result and the truth about the population due to practical difficulties such as contacting the subjects selected.
(d) any difference between the sample result and the truth about the population that tends to occur in the same direction whenever you use this sampling method.
(e) racism or sexism on the part of those who take the sample.

T4.9 You wonder if TV ads are more effective when they are longer or repeated more often or both. So you design an experiment. You prepare 30-second and 60-second ads for a camera. Your subjects all watch the same TV program, but you assign them at random to four groups. One group sees the 30-second ad once during the program; another sees it three times; the third group sees the 60-second ad once; and the last group sees the 60-second ad three times. You ask all subjects how likely they are to buy the camera.

(a) This is a randomized block design, but not a matched pairs design.
(b) This is a matched pairs design.
(c) This is a completely randomized design with one explanatory variable (factor).
(d) This is a completely randomized design with two explanatory variables (factors).
(e) This is a completely randomized design with four explanatory variables (factors).

T4.10 A researcher wishes to compare the effects of two fertilizers on the yield of soybeans. She has 20 plots of land available for the experiment, and she decides to use a matched pairs design with 10 pairs of plots. To carry out the random assignment for this design, the researcher should

(a) use a table of random numbers to divide the 20 plots into 10 pairs and then, for each pair, flip a coin to assign the fertilizers to the 2 plots.
(b) subjectively divide the 20 plots into 10 pairs (making the plots within a pair as similar as possible) and then, for each pair, flip a coin to assign the fertilizers to the 2 plots.
(c) use a table of random numbers to divide the 20 plots into 10 pairs and then use the table of random numbers a second time to decide on the fertilizer to be applied to each member of the pair.
(d) flip a coin to divide the 20 plots into 10 pairs and then, for each pair, use a table of random numbers to assign the fertilizers to the 2 plots.
(e) use a table of random numbers to assign the two fertilizers to the 20 plots and then use the table of random numbers a second time to place the plots into 10 pairs.

T4.11 You want to know the opinions of American high school teachers on the issue of establishing a national proficiency test as a prerequisite for graduation from high school. You obtain a list of all high school teachers belonging to the National Education Association (the country's largest teachers' union) and mail a survey to a random sample of 2500 teachers. In all, 1347 of the teachers return the survey. Of those who responded, 32% say that they favor some kind of national proficiency test. Which of the following statements about this situation is true?

(a) Because random sampling was used, we can feel confident that the percent of all American high school teachers who would say they favor a national proficiency test is close to 32%.
(b) We cannot trust these results, because the survey was mailed. Only survey results from face-to-face interviews are considered valid.
(c) Because over half of those who were mailed the survey actually responded, we can feel fairly confident that the actual percent of all American high school teachers who would say they favor a national proficiency test is close to 32%.
(d) The results of this survey may be affected by nonresponse bias.
(e) The results of this survey cannot be trusted due to voluntary response bias.

Section II: Free Response *Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.*

T4.12 Elephants sometimes damage trees in Africa. It turns out that elephants dislike bees. They recognize beehives in areas where they are common and avoid them. Can this be used to keep elephants away from trees? Will elephant damage be less in trees with hives? Will even empty hives keep elephants away? Researchers want to design an experiment to answer these questions using 72 acacia trees.[64]

(a) Identify the experimental units, treatments, and the response variable.
(b) Describe how the researchers could carry out a completely randomized design for this experiment. Include a description of how the treatments should be assigned.

T4.13 A *New York Times* article on public opinion about steroid use in baseball discussed the results of a sample survey. The survey found that 34% of adults think that at least half of Major League Baseball (MLB) players "use steroids to enhance their athletic performance." Another 36% thought that about a quarter of MLB players use steroids; 8% had no opinion. Here is part of the *Times*'s statement on "How the Poll Was Conducted":

> The latest New York Times/CBS News Poll is based on telephone interviews conducted March 15 through March 18 with 1,067 adults throughout the United States.… The sample of telephone numbers called was randomly selected by a computer from a list of more than 42,000 active residential exchanges across the country. The exchanges were chosen to ensure that each region of the country was represented in proportion to its population. In each exchange, random digits were added to form a complete telephone number, thus permitting access to listed and unlisted numbers. In each household, one adult was designated by a random procedure to be the respondent for the survey.[65]

(a) Explain why the sampling method used in this survey was *not* a simple random sample.

(b) Why was one adult chosen at random in each household to respond to the survey?

(c) Explain how undercoverage could lead to bias in this sample survey.

T4.14 Many people start their day with a jolt of caffeine from coffee or a soft drink. Most experts agree that people who take in large amounts of caffeine each day may suffer from physical withdrawal symptoms if they stop ingesting their usual amounts of caffeine. Researchers recruited 11 volunteers who were caffeine dependent and who were willing to take part in a caffeine withdrawal experiment. The experiment was conducted on two 2-day periods that occurred one week apart. During one of the 2-day periods, each subject was given a capsule containing the amount of caffeine normally ingested by that subject in one day. During the other study period, the subjects were given placebos. The order in which each subject received the two types of capsules was randomized. The subjects' diets were restricted during each of the study periods. At the end of each 2-day study period, subjects were evaluated using a tapping task in which they were instructed to press a button 200 times as fast as they could.[66]

(a) How and why was blocking used in the design of this experiment?

(b) Why did researchers randomize the order in which subjects received the two treatments?

(c) Could this experiment have been carried out in a double-blind manner? Explain.

Cumulative AP® Practice Test 1

Section I: Multiple Choice *Choose the best answer for Questions AP1.1 to AP1.14.*

AP1.1 You look at real estate ads for houses in Sarasota, Florida. Many houses range from $200,000 to $400,000 in price. The few houses on the water, however, have prices up to $15 million. Which of the following statements best describes the distribution of home prices in Sarasota?

(a) The distribution is most likely skewed to the left, and the mean is greater than the median.

(b) The distribution is most likely skewed to the left, and the mean is less than the median.

(c) The distribution is roughly symmetric with a few high outliers, and the mean is approximately equal to the median.

(d) The distribution is most likely skewed to the right, and the mean is greater than the median.

(e) The distribution is most likely skewed to the right, and the mean is less than the median.

AP1.2 A child is 40 inches tall, which places her at the 90th percentile of all children of similar age. The heights for children of this age form an approximately Normal distribution with a mean of 38 inches. Based on this information, what is the standard deviation of the heights of all children of this age?

(a) 0.20 inches (c) 0.65 inches (e) 1.56 inches

(b) 0.31 inches (d) 1.21 inches

AP1.3 A large set of test scores has mean 60 and standard deviation 18. If each score is doubled, and then 5 is subtracted from the result, the mean and standard deviation of the new scores are

(a) mean 115; std. dev. 31. (d) mean 120; std. dev. 31.

(b) mean 115; std. dev. 36. (e) mean 120; std. dev. 36.

(c) mean 120; std. dev. 6.

AP1.4 For a certain experiment, the available experimental units are eight rats, of which four are female (F1, F2, F3, F4) and four are male (M1, M2, M3, M4). There are to be four treatment groups, A, B, C, and D. If a randomized block design is used, with the experimental units blocked by gender, which of the following assignments of treatments is impossible?

(a) A → (F1, M1), B → (F2, M2), C → (F3, M3), D → (F4, M4)

(b) A → (F1, M2), B → (F2, M3), C → (F3, M4), D → (F4, M1)

(c) A → (F1, M2), B → (F3, F2), C → (F4, M1), D → (M3, M4)

(d) A → (F4, M1), B → (F2, M3), C → (F3, M2), D → (F1, M4)

(e) A → (F4, M1), B → (F1, M4), C → (F3, M2), D → (F2, M3)

AP1.5 For a biology project, you measure the weight in grams (g) and the tail length in millimeters (mm) of a group of mice. The equation of the least-squares line for predicting tail length from weight is

$$\text{predicted tail length} = 20 + 3 \times \text{weight}$$

Which of the following is *not* correct?

(a) The slope is 3, which indicates that a mouse's weight should increase by about 3 grams for each additional millimeter of tail length.

(b) The predicted tail length of a mouse that weighs 38 grams is 134 millimeters.

(c) By looking at the equation of the least-squares line, you can see that the correlation between weight and tail length is positive.

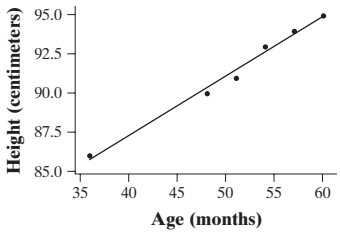
R3.5 (a) $\hat{y} = 30.2 + 0.16x$, where y = final exam score and x = total score before the final examination. (b) 78.2 (c) Of all the lines that the professor could use to summarize the relationship between final exam score and total points before the final exam, the least-squares regression line is the one that has the smallest sum of squared residuals. (d) Because $r^2 = 0.36$, only 36% of the variability in the final exam scores is accounted for by the linear model relating final exam scores to total score before the final exam. More than half (64%) of the variation in final exam scores is *not* accounted for, so Julie has reason to question this estimate.

R3.6 Even though there is a high correlation between number of calculators and math achievement, we shouldn't conclude that increasing the number of calculators will *cause* an increase in math achievement. It is possible that students who are more serious about school have better math achievement and also have more calculators.

Answers to Chapter 3 AP® Statistics Practice Test

T3.1 d
T3.2 e
T3.3 c
T3.4 a
T3.5 a
T3.6 c
T3.7 b
T3.8 e
T3.9 b
T3.10 c

T3.11 (a) A scatterplot with regression line is shown below. (b) $\hat{y} = 71.95 + 0.3833x$, where y = height and x = age. (c) 255.934 cm, or 100.76 inches (d) This was an extrapolation. Our data were based only on the first 5 years of life and the linear trend will not continue forever.



T3.12 (a) The point in the upper-right-hand corner has a very high silicon value for its isotope value. (b) (i) r would get closer to -1 because it does not follow the linear pattern of the other points. (ii) Because this point is "pulling up" the line on the right side of the plot, removing it will make the slope steeper (more negative) and the y intercept smaller (note that the y axis is to the *right* of the points in the scatterplot). (iii) Because this point has a large residual, removing it will make s a little smaller.

T3.13 (a) $\hat{y} = 92.29 - 0.05762x$, where y is the percent of the grass burned and x is the number of wildebeest. (b) For every increase of 1000 wildebeest, the predicted percent of grassy area burned decreases by about 0.058. (c) $r = -\sqrt{0.646} = -0.804$. There is a strong, negative linear association between the percent of grass burned and the number of wildebeest. (d) Yes, because there is no obvious leftover pattern in the residual plot.

Chapter 4

Section 4.1

Answers to Check Your Understanding

page 213: **1.** Convenience sampling. This could lead the inspector to overestimate the quality of the oranges if the farmer puts the best oranges on top. **2.** Voluntary response sampling. In this case, those who are happy that the UN has its headquarters in the U.S. already have what they want and so are less likely to respond. The proportion who answered "No" in the sample is likely to be higher than the true proportion in the U.S. who would answer "No."

page 223: **1.** You would have to identify 200 different seats, go to those seats in the arena, and find the people who are sitting there, which would take a lot of time. **2.** It is best to create strata where the people within a stratum are very similar to each other but different than the people in other strata. In this case, it would be better to take the lettered rows as the strata because each lettered row is the same distance from the court and so would contain only seats with the same (or nearly the same) ticket price. **3.** It is best if the people in each cluster reflect the variability found in the population. In this case, it would be better to take the numbered sections as the clusters because they include all different seat prices.

page 228: **1.** (a) Undercoverage (b) Nonresponse (c) Undercoverage **2.** By making it sound like they are not a problem in the landfill, this question will result in fewer people suggesting that we should ban disposable diapers. The proportion who would say "Yes" to this survey question is likely to be smaller than the proportion who would say "Yes" to a more fairly worded question.

Answers to Odd-Numbered Section 4.1 Exercises

4.1 Population: all local businesses. Sample: the 73 businesses that return the questionnaire.

4.3 Population: the 1000 envelopes stuffed during a given hour. Sample: the 40 randomly selected envelopes.

4.5 This is a voluntary response sample. In this case, it appears that people who strongly support gun control volunteered more often, causing the proportion in the sample to be greater than the proportion in the population.

4.7 This is a voluntary response sample and overrepresents the opinions of those who feel most strongly about the issue being surveyed.

4.9 (a) A convenience sample (b) The first 100 students to arrive at school likely had to wake up earlier than other students, so 7.2 hours is probably less than the true average.

4.11 (a) Number the 40 students from 01 to 40. Pick a starting point on the random number table. Record two-digit numbers, skipping numbers that aren't between 01 and 40 and any repeated numbers, until you have 5 unique numbers between 01 and 40. Use the 5 students corresponding to these numbers. (b) Using line 107, skip the numbers not in bold: 82 73 95 78 90 **20** 80 74 75 **11** 81 67 65 53 00 94 **38** **31** 48 93 60 94 **07**. Select Johnson (20), Drasin (11), Washburn (38), Rider (31), and Calloway (07).

4.13 (a) *Using calculator*: Number the plots from 1 to 1410. Use the command `randInt(1,1410)` to select 141 different integers from 1 to 1410 and use the corresponding 141 plots. (b) Answers will vary.

4.15 (a) False—although, on average, there will be four 0s in every set of 40 digits, the number of 0s can be less than 4 or greater than 4 by chance. (b) True—there are 100 pairs of digits 00 through 99,

and all are equally likely. (c) False—0000 is just as likely as any other string of four digits.

4.17 (a) It might be difficult to locate the 20 phones from among the 1000 produced that day. (b) The quality of the phones produced may change during the day, so that the last phones manufactured are not representative of the day's production. (c) Because each sample of 20 phones does not have the same probability of being selected. In an SRS, it is possible for 2 consecutive phones to be selected in a sample, but this is not possible with a systematic random sample.

4.19 Assign numbers 01 to 30 to the students. Pick a starting point on the random digit table. Record two-digit numbers, skipping any that aren't between 01 and 30 and any repeated numbers, until you have 4 unique numbers between 01 and 30. Use the corresponding four students. Then assign numbers 0 to 9 to the faculty members. Continuing on the table, record one-digit numbers, skipping any repeated numbers, until you have 2 unique numbers between 0 and 9. Use the corresponding faculty members. Starting on line 123 gives 08-Ghosh, 15-Jones, 07-Fisher, and 27-Shaw for the students and 1-Besicovitch and 0-Andrews for the faculty.

4.21 (a) Use the three types of seats as the strata because people who can afford more expensive tickets probably have different opinions about the concessions than people who can afford only the cheaper tickets. (b) A stratified random sample will include seats from all over the stadium, which would make it very time-consuming to obtain. A cluster sample of numbered sections would be easier to obtain, because the people selected for the sample would be sitting close together.

4.23 No. In an SRS, each possible sample of 250 engineers is equally likely to be selected, including samples that aren't exactly 200 males and 50 females.

4.25 (a) Cluster sampling. (b) To save time and money. In an SRS, the company would have to visit individual homes all over the rural subdivision instead of only 5 locations.

4.27 (a) It is unlikely, because different random samples will include different students and produce different estimates of the proportion of students who use Twitter. (b) An SRS of 100 students. Larger random samples give us better information about the population than smaller random samples.

4.29 Because you are sampling only from the lower-priced ticket holders, this will likely produce an estimate that is too small, as fans in the club seats and box seats probably spend more money at the game than fans in cheaper seats.

4.31 (a) 89.1% (b) Because the people who have long commutes are less likely to be at home and be included in the sample, this will likely produce an estimate that is too small.

4.33 We would not expect very many people to claim they have run red lights when they haven't, but some people will deny running red lights when they have. Thus, we expect that the sample proportion underestimates the true proportion of drivers who have run a red light.

4.35 (a) The wording is clear, but the question is slanted in favor of warning labels because of the first sentence stating that some cell phone users have developed brain cancer. (b) The question is clear, but it is slanted in favor of national health insurance by asserting it would reduce administrative costs and not providing any counterarguments. (c) The wording is too technical for many people to understand. For those who do understand the question, it is slanted because it suggests reasons why one should support recycling.

4.37 c

4.39 d

4.41 d

4.43 (a) For each additional day, the predicted sleep debt increases by about 3.17 hours. (b) The predicted sleep debt for a 5-day school week is $2.23 + 3.17(5) = 18.08$ hours. This is about 3 hours more than the researcher claimed for a 5-day week, so the students have reason to be skeptical of the research study's reported results.

Section 4.2

Answers to Check Your Understanding

page 237: **1.** Experiment, because a treatment (brightness of screen) was imposed on the laptops. **2.** Observational study, because students were not assigned to eat a particular number of meals with their family per week. **3.** Explanatory: number of meals per week eaten with their family. Response: GPA. **4.** There are probably other variables that are influencing the response variable. For example, students who have part-time jobs may not be able to eat many meals with their families and may not have much time to study, leading to lower grades.

page 247: **1.** Randomly assign the 29 students to two treatments: evaluating the performance in small groups or evaluating the performance alone. The response variable will be the accuracy of their final performance evaluations. To implement this design, use 29 equally sized slips of paper. Label 15 of them "small group" and 14 of them "alone." Then shuffle the papers and hand them out at random to the 29 students, assigning them to a treatment. **2.** The purpose of the control group is to provide a baseline for comparison. Without a group to compare to, it is impossible to determine if the small group treatment is more effective.

page 249: **1.** No. Perhaps seeing the image of their unborn child encouraged the mothers who had an ultrasound to eat a better diet, resulting in healthier babies. **2.** No. While the people weighing the babies at birth may not have known whether that particular mother had an ultrasound or not, the mothers knew. This might have affected the outcome because the mothers knew whether they had received the treatment or not. **3.** Treat all mothers as if they had an ultrasound, but for some mothers the ultrasound machine wouldn't be turned on. To avoid having mothers know the machine was turned off, the ultrasound screen would have to be turned away from all the mothers.

Answers to Odd-Numbered Section 4.2 Exercises

4.45 Experiment, because students were randomly assigned to the different teaching methods.

4.47 (a) Observational study, because mothers weren't assigned to eat different amounts of chocolate. (b) Explanatory: the mother's chocolate consumption. Response: the baby's temperament. (c) No, this study is an observational study so we cannot draw a cause-and-effect conclusion. It is possible that women who eat chocolate daily have less stressful lives and the lack of stress helps their babies to have better temperaments.

4.49 Type of school. For example, private schools tend to have smaller class sizes and students that come from families with higher socioeconomic status. If these students do better in the future, we wouldn't know if the better performance was due to smaller class sizes or higher socioeconomic status.

4.51 Experimental units: pine seedlings. Explanatory variable: light intensity. Response variable: dry weight at the end of the study. Treatments: full light, 25% light, and 5% light.

4.53 Experimental units: the individuals who were called. Explanatory variables: (1) information provided by interviewer; (2) whether caller offered survey results. Response variable: whether or not the call was completed. Treatments: (1) name/no offer; (2) university/no offer; (3) name and university/no offer; (4) name/offer; (5) university/offer; (6) name and university/offer.

4.55 Experimental units: 24 fabric specimens. Explanatory variables: (1) roller type; (2) dyeing cycle time; (3) temperature. Response variable: a quality score. Treatments: (1) metal, 30 min, 150°; (2) natural, 30 min, 150°; (3) metal, 40 min, 150°; (4) natural, 40 min, 150°; (5) metal, 30 min, 175°; (6) natural, 30 min, 175°; (7) metal, 40 min, 175°; (8) natural, 40 min, 175°.

4.57 There was no control group. We don't know if the improvement was due to the placebo effect or if the flavonols actually affected the blood flow.

4.59 **(a)** Write all names on slips of paper, put them in a container, and mix thoroughly. Pull out 40 slips of paper and assign these subjects to Treatment 1. Then pull out 40 more slips of paper and assign these subjects to Treatment 2. The remaining 40 subjects are assigned to Treatment 3. **(b)** Assign the students numbers from 1 to 120. Using the command `RandInt (1,120)` on the calculator, assign the students corresponding to the first 40 unique numbers chosen to Treatment 1, the students corresponding to the next 40 unique numbers chosen to Treatment 2, and the remaining 40 students to Treatment 3. **(c)** Assign the students numbers from 001 to 120. Pick a spot on Table D and read off the first 40 unique numbers between 001 and 120. The students corresponding to these numbers are assigned to Treatment 1. The students corresponding to the next 40 unique numbers between 001 and 120 are assigned to Treatment 2. The remaining 40 students are assigned to Treatment 3.

4.61 Random assignment. If players are allowed to choose which treatment they get, perhaps the more motivated players will choose the new method. If they improve more by the end of the study, the coach can't be sure if it was the exercise program or player motivation that caused the improvement.

4.63 *Comparison:* Researchers used a design that compared a low-carbohydrate diet with a low-fat diet. *Random assignment:* Subjects were randomly assigned to one of the two diets. *Control:* The experiment used subjects who were all obese at the beginning of the study and who all lived in the same area. *Replication:* There were 66 subjects in each treatment group.

4.65 Write the names of the patients on 36 identical slips of paper, put them in a hat, and mix them well. Draw out 9 slips. The corresponding patients will receive the antidepressant. Draw out 9 more slips. Those patients will receive the antidepressant plus stress management. The patients corresponding to the next 9 slips drawn will receive the placebo, and the remaining 9 patients will receive the placebo plus stress management. At the end of the experiment, record the number and severity of chronic tension-type headaches for each of the 36 subjects and compare the results for the 4 groups.

4.67 **(a)** Other variables include expense and condition of the patient. For example, if a patient is in very poor health, a doctor might choose not to recommend surgery because of the added complications. Then we won't know if a higher death rate is due to the treatment or the initial health of the subjects. **(b)** Write the names of all 300 patients on identical slips of paper, put them in a hat, and mix them well. Draw out 150 slips and assign the corresponding subjects to receive surgery. The remaining 150 subjects receive the new method. At the end of the study, count how many patients survived in each group.

4.69 The subjects developed rashes on the arm exposed to the placebo (a harmless leaf) simply because they thought they were being exposed to a poison ivy leaf. Likewise, most of the subjects didn't develop rashes on the arm that was exposed to poison ivy because they didn't think they were being exposed to the real thing.

4.71 Because the experimenter knew which subjects had learned the meditation techniques, he is not blind. If the experimenter believed that meditation was beneficial, he may subconsciously rate subjects in the meditation group as being less anxious.

4.73 **(a)** To make sure that the two groups were as similar as possible before the treatments were administered. **(b)** The difference in weight loss was larger than would be expected due to the chance variation created by the random assignment to treatments. **(c)** Even though the low-carb dieters lost 2 kg more over the year than the low-fat group, a difference of 2 kg could be due just to chance variation created by the random assignment.

4.75 **(a)** The different diagnoses, because the treatments were randomly assigned to patients within each diagnosis. **(b)** Using a randomized block design allows us to account for the variability in response due to differences in diagnosis by initially comparing the results within each block. In a completely randomized design, this variability will be unaccounted for, making it harder to determine if there is a difference in health and satisfaction due to the difference between doctors and nurse-practitioners.

4.77 **(a)** A randomized block design would help us account for the variability in yield that is due to the differences in fertility in the field, making it easier to determine if one variety is better than the others. **(b)** The rows. There should be a stronger association between row number and yield than column number and yield. **(c)** Let the digits 1 to 5 correspond to the five corn varieties A to E. Begin with line 111 on the random digit table, and assign the letters to the top row from left to right, ignoring numbers 0 and 6–9 and repeated numbers. Use a different line (111, 112, 113, 114, and 115) for each row. Top row (left to right): ADECB, second row: ECDAB, third row: BEDCA, fourth row: DEACB, bottom row: ADCBE.

4.79 **(a)** If all rats from litter 1 were fed Diet A and if these rats gained more weight, we would not know if this was because of the diet or because of genetics and initial health. **(b)** Use a randomized block design with the litters as blocks. For each of the litters, randomly assign half of the rats to receive Diet A and the other half to receive Diet B. This will allow researchers to account for the differences in weight gain caused by the differences in genetics and initial health.

4.81 **(a)** Matched pairs design. **(b)** In a completely randomized design, the differences between the students will add variability to the response, making it harder to detect if there is a difference caused by the treatments. In a matched pairs design, each student is compared with himself (or herself), so the differences between students are accounted for. **(c)** If all the students used the hands-free phone during the first session and performed worse, we wouldn't know if the better performance during the second session is due to the lack of phone or to learning from their mistakes the first time. By randomizing the order, some students will use the hands-free phone during the first session and others during the second session. **(d)** The simulator, route, driving conditions, and traffic flow were all kept the same for both sessions, preventing these variables from adding variability to the response variable.

4.83 **(a)** Randomly assign the 20 subjects into two groups of 10. Write the name of each subject on a note card, shuffle the cards, and select 10 to be assigned to the 70° environment. The remaining 10 subjects will be assigned to the 90° environment. Then the number of correct insertions will be recorded for each subject and the two groups compared. **(b)** All subjects will perform the task twice, once in each temperature condition. Randomly choose the order by flipping a coin. Heads: 70°, then 90°. Tails: 90°, then 70°. For each subject, compare the number of correct insertions in each environment.

4.85 **(a)** If the students find a difference between the two groups, they will not know if the difference is due to gender or the deodorant. **(b)** Each student should have one armpit randomly assigned to receive Deodorant A and the other Deodorant B. Because each gender uses both deodorants, there is no longer any confounding between gender and deodorant.

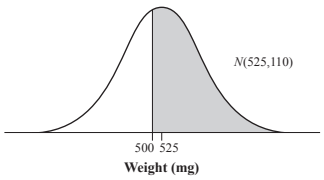
4.87 c

4.89 b

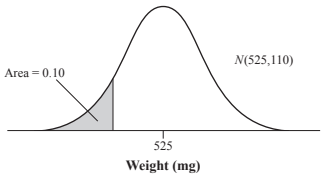
4.91 c

4.93 b

4.95 **(a)** For these seeds, the weights follow a $N(525, 110)$ distribution and we want the proportion of seeds that weigh more than 500 mg (see graph below). $z = \frac{500 - 525}{110} = -0.23$. From Table A, the proportion of z-scores greater than -0.23 is $1 - 0.4090 = 0.5910$. *Using technology* `normalcdf(lower:500,upper:10000, μ:525,σ:110)` $= 0.5899$. About 59% of seeds will weigh more than 500 mg.



(b) For these seeds, the weights follow a $N(525, 110)$ distribution and we are looking for the boundary value x that has an area of 0.10 to the left (see graph below). A z-score of -1.28 gives the closest value to 0.10 (0.1003). Solving $-1.28 = \frac{x - 525}{110}$ gives $x = 384.2$. *Using technology:* `invNorm(area:0.10,μ:525,σ:110)` $=$ 384.0. The smallest weight among the remaining seeds should be about 384 mg.



Section 4.3

Answers to Odd-Numbered Section 4.3 Exercises

4.97 If the study involves random sampling, we can make inferences about the population from which we sampled. If the study involves random assignment, we can make inferences about cause and effect.

4.99 Because this study involved random assignment to the treatments, we can infer that the difference between foster care or institutional care caused the difference in response.

4.101 Because this study did not involve random assignment to a treatment, we cannot infer cause and effect. Also, because the individuals were not randomly chosen, we cannot generalize to a larger population.

4.103 As daytime running lights become more common, they may be less effective at catching the attention of other drivers. Also, a driving simulator might not be very realistic.

4.105 Answers will vary.

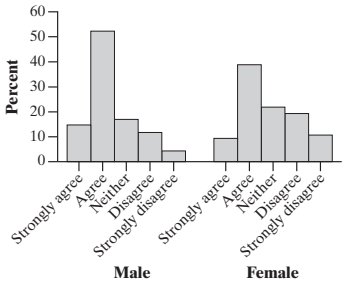
4.107 Answers will vary.

4.109 Confidential. The person taking the survey knows who is answering the questions, but will not share the results of individuals with anyone else.

4.111 The subjects were not able to give informed consent. They did not know what was happening to them and they were not old enough to understand the ramifications.

4.113 The conditional distributions for males and females are displayed in the table and graph below. Men are more likely to view animal testing as justified if it might save human lives: over two-thirds of men agree or strongly agree with this statement, compared to slightly less than half of the women. The percentages who disagree or strongly disagree tell a similar story: 16% of men versus 30% of women.

Response	Male	Female
Strongly agree	14.7%	9.3%
Agree	52.3%	38.8%
Neither	16.9%	21.9%
Disagree	11.8%	19.3%
Strongly disagree	4.3%	10.7%



Answers to Chapter 4 Review Exercises

R4.1 **(a)** Population: all Ontario residents. Sample: the 61,239 people interviewed. **(b)** Because different samples will produce different estimates, it is unlikely that the percentages in the entire population would be exactly the same as the percentages in the sample. However, they should be fairly close.

R4.2 **(a)** Announce in a daily bulletin that there is a survey concerning student parking available in the main office for students who want to respond. Because those who feel strongly are more likely to respond, their opinions will be overrepresented. **(b)** Interview a group of students as they come in from the parking lot. People who already can park on campus might have different opinions about the parking situation than those who cannot.

R4.3 (**a**) Number the players from 01 to 25 in alphabetical order. Move from left to right, reading pairs of digits until you find three different pairs between 01 and 25, and select the corresponding players. (**b**) 17 (Musselman), 09 (Fuhrmann), and 23 (Smith).

R4.4 Stratified, because it is likely that the opinions of professors will vary based on which type of institution they are at. Then a stratified random sample will provide a more precise estimate than the other methods. Furthermore, the other methods might miss faculty from one particular type of institution.

R4.5 (**a**) People may not remember how many movies they watched in a movie theater in the past year. So shorten the amount of time that they ask about, perhaps 3 or 6 months. (**b**) This will underrepresent younger adults who use only cell phones. If younger adults go to movies more often than older adults, the estimated mean will be too small. (**c**) Because the frequent moviegoers will not be at home to respond, the estimated mean will be too small.

R4.6 (**a**) Different anesthetics were not randomly assigned to the subjects. (**b**) Type of surgery. If Anesthesia C is used more often with a type of surgery that has a higher death rate, we wouldn't know if the death rate was higher because of the anesthesia or the type of surgery.

R4.7 (**a**) Units: potatoes. Explanatory: storage method and time from slicing until cooking. Response: ratings of color and flavor. Treatments: (1) fresh/immediately, (2) fresh/after an hour, (3) room temperature/immediately, (4) room temperature/after an hour, (5) refrigerator/immediately, (6) refrigerator/after an hour. (**b**) Using 300 identical slips of paper, write "1" on 50 of them, "2" on 50 of them, and so on. Put the papers in a hat and mix well. Then select a potato and randomly select a slip from the hat to determine which treatment that potato will receive. Repeat this process for the remaining 299 potatoes, making sure not to replace the slips of paper into the hat. (**c**) Use a randomized block design with regular potatoes in one block and sweet potatoes in the other block. Randomly assign the 6 treatments within each block as in part (b).

R4.8 (**a**) No. The 1000 students were not randomly selected from any larger population. (**b**) Yes. The students were randomly assigned to the three treatments.

R4.9 (**a**) By giving some patients a treatment that should have no effect at all, but appears like the Saint-John's-wort, the researchers can account for the expectations of patients (the placebo effect) by comparing the results for the two groups. (**b**) To create two groups of subjects that are roughly equivalent at the beginning of the experiment. (**c**) The subjects should not know which treatment they are getting so that the researchers can account for the placebo effect. The researchers should be unaware of which subjects received which treatment so that they cannot influence how the results are measured. (**d**) The difference in improvement between the two groups wasn't large enough to rule out the chance variation caused by the random assignment to treatments.

R4.10 (**a**) Randomly assign 15 students to easy mazes and the other 15 to hard mazes. Use 30 identical slips of paper and write the name of each subject on a slip. Mix the slips in a hat, select 15 of them at random, and assign these subjects to hard mazes. The remaining 15 will be assigned to easy mazes. After the experiment, compare the time estimates of the two groups. (**b**) Each student does the activity twice, once with each type of maze. Randomly determine which set of mazes is used first by flipping a coin for each subject. Heads: easy, then hard. Tails: hard, then easy. After the experiment, compare each student's easy maze and hard maze time estimate. (**c**) The matched pairs design would be more likely to detect a difference because it accounts for the variability between subjects.

R4.11 (**a**) This does not meet the requirements of informed consent because the subjects did not know the nature of the experiment before they agreed to participate. (**b**) All individual data should be kept confidential and the experiment should go before an institutional review board before being implemented.

Answers to Chapter 4 AP® Statistics Practice Test

T4.1 c
T4.2 e
T4.3 d
T4.4 c
T4.5 b
T4.6 b
T4.7 d
T4.8 d
T4.9 d
T4.10 b
T4.11 d

T4.12 (**a**) Experimental units: acacia trees. Treatments: placing either active beehives, empty beehives, or nothing in the trees. Response: damage to the trees caused by elephants. (**b**) Assign the trees numbers from 01 to 72 and use a random number table to pick 24 different two-digit numbers in this range. Those trees will get the active beehives. The trees corresponding to the next 24 different two-digit numbers from 01 to 72 will get the empty beehives, and the remaining 24 trees will remain empty. Compare the damage caused by elephants to the three groups of trees.

T4.13 (**a**) Not all possible samples of size 1067 were possible. For example, using their method, they could not have had all respondents from the east coast. (**b**) If the household members who typically answer the phone have a different opinion than those who don't typically answer the phone, their opinions will be overrepresented. (**c**) If people without phones or with cell phones only have different opinions than the group of people with residential lines, these opinions will be underrepresented.

T4.14 (**a**) Each of the 11 individuals will be a block in this matched pairs design, with the order of treatments randomly assigned. This was to help account for the variability in tapping speed caused by the differences in subjects. (**b**) If all the subjects got caffeine the second time, the researchers wouldn't know if the increase was due to the caffeine or due to practice with the task. (**c**) Yes. Neither the subjects nor the people who come in contact with them during the experiment (including those who record the number of taps) need to know the order in which the caffeine or placebo was administered.

Answers to Cumulative AP® Practice Test 1

AP1.1 d
AP1.2 e
AP1.3 b
AP1.4 c
AP1.5 a
AP1.6 c
AP1.7 e
AP1.8 e
AP1.9 d
AP1.10 d
AP1.11 d

Table D Random digits

Line								
101	19223	95034	05756	28713	96409	12531	42544	82853
102	73676	47150	99400	01927	27754	42648	82425	36290
103	45467	71709	77558	00095	32863	29485	82226	90056
104	52711	38889	93074	60227	40011	85848	48767	52573
105	95592	94007	69971	91481	60779	53791	17297	59335
106	68417	35013	15529	72765	85089	57067	50211	47487
107	82739	57890	20807	47511	81676	55300	94383	14893
108	60940	72024	17868	24943	61790	90656	87964	18883
109	36009	19365	15412	39638	85453	46816	83485	41979
110	38448	48789	18338	24697	39364	42006	76688	08708
111	81486	69487	60513	09297	00412	71238	27649	39950
112	59636	88804	04634	71197	19352	73089	84898	45785
113	62568	70206	40325	03699	71080	22553	11486	11776
114	45149	32992	75730	66280	03819	56202	02938	70915
115	61041	77684	94322	24709	73698	14526	31893	32592
116	14459	26056	31424	80371	65103	62253	50490	61181
117	38167	98532	62183	70632	23417	26185	41448	75532
118	73190	32533	04470	29669	84407	90785	65956	86382
119	95857	07118	87664	92099	58806	66979	98624	84826
120	35476	55972	39421	65850	04266	35435	43742	11937
121	71487	09984	29077	14863	61683	47052	62224	51025
122	13873	81598	95052	90908	73592	75186	87136	95761
123	54580	81507	27102	56027	55892	33063	41842	81868
124	71035	09001	43367	49497	72719	96758	27611	91596
125	96746	12149	37823	71868	18442	35119	62103	39244
126	96927	19931	36809	74192	77567	88741	48409	41903
127	43909	99477	25330	64359	40085	16925	85117	36071
128	15689	14227	06565	14374	13352	49367	81982	87209
129	36759	58984	68288	22913	18638	54303	00795	08727
130	69051	64817	87174	09517	84534	06489	87201	97245
131	05007	16632	81194	14873	04197	85576	45195	96565
132	68732	55259	84292	08796	43165	93739	31685	97150
133	45740	41807	65561	33302	07051	93623	18132	09547
134	27816	78416	18329	21337	35213	37741	04312	68508
135	66925	55658	39100	78458	11206	19876	87151	31260
136	08421	44753	77377	28744	75592	08563	79140	92454
137	53645	66812	61421	47836	12609	15373	98481	14592
138	66831	68908	40772	21558	47781	33586	79177	06928
139	55588	99404	70708	41098	43563	56934	48394	51719
140	12975	13258	13048	45144	72321	81940	00360	02428
141	96767	35964	23822	96012	94591	65194	50842	53372
142	72829	50232	97892	63408	77919	44575	24870	04178
143	88565	42628	17797	49376	61762	16953	88604	12724
144	62964	88145	83083	69453	46109	59505	69680	00900
145	19687	12633	57857	95806	09931	02150	43163	58636
146	37609	59057	66967	83401	60705	02384	90597	93600
147	54973	86278	88737	74351	47500	84552	19909	67181
148	00694	05977	19664	65441	20903	62371	22725	53340
149	71546	05233	53946	68743	72460	27601	45403	88692
150	07511	88915	41267	16853	84569	79367	32337	03316